

Event predictability modulates production and comprehension of referring expressions

Ivan Rygaev^a, Vera Demberg^b, Martin V. Butz^c, Asya Achimova^a

*^aDepartment of Linguistics, University of Tübingen,
Keplerstraße 2, Tübingen, 72074, Germany*

*^bDepartment of Computer Science and Department of Language Science and Technology,
Saarland University, Campus C7.2, Saarbrücken, 66123, Germany*

*^cCognitive Modeling, Department of Computer Science and Department of Psychology,
University of Tübingen, Sand 14, Tübingen, 72076, Germany*

Abstract

When speakers choose how to refer to a previously mentioned character in the presence of several potential antecedents, they experience multiple competing pressures. Predictability is an important factor in this process, and its role in shaping speakers' choices has been the topic of ongoing debate. Here, we examine how event predictability, viewed as a specific form of information predictability, shapes both the production and interpretation of anaphoric references. In Experiments 1a and 1b (free production), participants described events involving interacting characters. They used overt forms of reference, such as definite noun phrases, more often when describing surprising events compared to familiar ones. Speakers further produced fewer phonologically non-overt forms, such as zero anaphors, when events were surprising. This pattern is consistent with sensitivity to possible listener misinterpretation of non-overt forms despite them being grammatically unambiguous. In Experiments 2a and 2b (speech perception), listeners heard descriptions of the same kinds of events under varying levels of auditory noise. Listeners were more likely to interpret utterances containing unambiguous zero anaphors as if they contained ambiguous pronouns when this interpretation allowed them to construe a more plausible event, particularly under noisy conditions. Taken together, the results show that event-level expectations shape both referential production and interpretation: speakers favor more overt forms of reference for less predictable events, whereas listeners use event plausibility when interpreting potentially degraded input. This convergence between production and comprehension is consistent with audience-design accounts of reference and extends noisy-channel accounts of communication to the domain of anaphoric interpretation.

Keywords: referring expressions, anaphora, background beliefs, events, predictability

1. Introduction

Anaphoric reference plays a major part in the process of establishing discourse coherence. When choosing forms of reference, speakers must balance communicative efficiency with referential clarity (Aylett and Turk, 2004; Levy and Jäger, 2007; Zipf, 1949). This balance requires speakers to choose expressions informative enough for listeners to identify the intended referent, while conserving articulatory effort via shorter, albeit potentially ambiguous forms. For example, when telling a short story about two monsters, one red and one yellow, as in (1), the speaker may choose to refer to a previously mentioned monster with a pronoun in (1b) rather than a definite noun phrase (NP) in (1a).

- (1) The yellow monster attacked the red monster and ...
- | | | |
|----|-----------------------------------|-------------|
| a. | the red monster fell down. | definite NP |
| b. | it fell down. | pronoun |

The question of what factors determine the likelihood of pronominalization in discourse has been a topic of ongoing debate, and predictability has taken a prominent place in this discussion (Davies and Arnold, 2019). A range of linguistic factors have been argued to affect referent predictability. They have been studied both through their effect on how speakers produce referring expressions and on how listeners resolve them. These include the grammatical function of the antecedent (Arnold, 1998; Grosz et al., 1995), thematic role (Kaiser et al., 2011), topicality (Rohde and Kehler, 2014), order of mention, and parallelism (Caramazza et al., 1977). Critically, these two lines of work have largely proceeded independently, and do not always converge on the same conclusions about which factors matter and how strongly. Predictable referents have been hypothesized to be more likely to be pronominalized, because listeners can more easily recover the intended antecedents (Arnold, 2001, 2008). Bott and Solstad (2023) refer to this link between next-mention likelihood and the form of the referring expression as the association hypothesis, and offer experimental evidence that the production of referring expressions is influenced by the likelihood of next mention.

In addition to the linguistic factors that shape referent predictability, the choice of referring expressions could be affected by the predictability of events in which these referents participate. The Gricean view of communication suggests that if a referent is not predictable, speakers should avoid ambiguous forms, such as pronouns, in referring to it, since an ambiguous reference can be misinterpreted. In that case, speakers are expected to choose more overt albeit more costly forms (Grice, 1975). In the event description that we considered in (1), the antecedent of the pronoun is easily recoverable, since our world knowledge suggests that if

one participant attacked another, it would be the victim of the attack who is more likely to fall down. On the other hand, a minimally different event in which the attacker itself falls down is less expected. Thus in (2), a speaker may prefer a definite NP in (2a) as a form of reference, rather than the pronoun in (2b).

- (2) The yellow monster attacked the red monster and ...
- | | | |
|----|--------------------------------------|--------------|
| a. | the yellow monster fell down. | definite NP |
| b. | it fell down. | pronoun |
| c. | fell down. | zero anaphor |

Additionally, if an event is surprising, a rational speaker may prefer to avoid a form with verb phrase (VP) coordination, like in (2c), where the subject of the second event is not explicitly mentioned. Even though VP coordination allows the speaker to avoid the potential ambiguity of the pronoun in (2b), a listener may interpret the utterance as though an overt pronoun were present, particularly when the speech signal is degraded by auditory noise.

In this paper, we investigate how event predictability interacts with linguistic factors in shaping the choice and interpretation of referring expressions. In Experiments 1a and 1b, we examine how speakers' choice of referring expression depends on event predictability. In Experiments 2a and 2b, we test the complementary prediction that listeners use event plausibility when interpreting referring expressions with and without added auditory noise. We argue for a view of reference production and comprehension as deeply intertwined with world knowledge: speakers and listeners do not merely track linguistic probabilities but also take into account the predictability of the events being described.

2. Background

Several distinct lines of work have examined how predictability affects speakers' choices. These lines of work differ in the experimental paradigms they employ, and, more critically, in how predictability is operationalized. In this section, we first review studies that treat predictability at the level of linguistic descriptions and then turn to those that focus on information predictability and its influence on language production. We then define event predictability as a particular kind of information predictability and consider its influence on the choice of anaphoric references in descriptions of events. We further draw upon two theoretical frameworks—the Rational Speech Act model and the noisy-channel hypothesis—that offer a computational basis for explaining the role of event predictability in speech production and comprehension. Finally, we contextualize our current study in relation to our prior work, which offers a distinct experimental paradigm to study the influence of event predictability on utterance choice and comprehension, separated from linguistic predictability.

2.1. Predictability of referents in discourse

Predictability has been argued to affect speakers’ choices at multiple levels of linguistic analysis. For example, speakers have been shown to mark unexpected referents by choosing acoustically prominent forms (Lam and Watson, 2010), using optional case marking more often (Kurumada and Jäger, 2015), and opting for more explicit syntactic structures (Jäger, 2010). In the domain of referring expressions, predictability has been commonly operationalized as the likelihood that the referent will be mentioned again in discourse. The choice of the next referent and the form of a referring expression have been related to the goal of establishing coherence relations between the parts of a sentence or discourse.

Prior work on the role of predictability has largely focused on the phenomenon of implicit causality—a semantic bias associated with a verb that shapes expectations about which participant will be mentioned in a subsequent causal explanation (Bott and Solstad, 2023; Caramazza et al., 1977; Demberg et al., 2023; Garvey and Caramazza, 1974; Garvey et al., 1974; Rohde and Kehler, 2014; Stevenson et al., 1994). To investigate the effect of predictability, these studies employed a passage completion task in which participants are required to continue a discourse passage with an (ambiguous) pronoun present in the prompt or without it, either with the instruction to talk about a specific referent or without such an instruction. Using this paradigm, Arnold (2001) demonstrated that the thematic role of the antecedent affects the likelihood of pronominalization, with goals being pronominalized more often than sources. Other studies have provided further evidence that the predictability of the referent affects the choice of referring expression. For example, Vogels (2019) showed the effect of next-mention bias independent of thematic role on the produced referring expressions in Dutch.

However, several other studies reported no effect of predictability on pronominalization (Fukumura and Van Gompel, 2010; Holler and Suckow, 2016; Modi et al., 2017; Rohde and Kehler, 2014). Kehler et al. (2008) found that while semantic cues, and specifically coherence relations, affect the likelihood of a particular referent being mentioned next, they do not determine the form of anaphoric reference. Thus, these studies argue for a dissociation between the next-mention likelihood and the form of the referring expression (Bott and Solstad, 2023). More recent studies have attributed the lack of effect of predictability on the choice of referring expressions in the aforementioned studies to design choices. Some authors have emphasized the role of the speaker’s engagement in the task (Demberg et al., 2023; Medina Fetterman et al., 2022; Rosa and Arnold, 2017; Weatherford and Arnold, 2021; Ye and Arnold, 2025; Zerkle and Arnold, 2019). In particular, Demberg et al. (2023) found that in an interactive setup in which participants are provided with more context, the effect of predictability was more pronounced.

Another feature that differentiates between studies that attested and those

that did not attest the role of predictability relates to potential ambiguity of the reference that speakers are required to produce. Arnold (2008) and Rohde and Kehler (2014) drew a distinction between studies where misinterpretation is at stake due to antecedents sharing properties, such as gender. When antecedents share gender, the use of pronouns can lead to ambiguity and potential misinterpretation, while designs with differently gendered antecedents substantially reduce this source of ambiguity. On the other hand, Arnold and Griffin (2007) showed that even the presence of a character of a different gender leads to reduced production of pronouns compared to the absence of a second character, even though the pronoun would be unambiguous in this case. In sum, the study of the effect of predictability on pronominalization requires controlling for the role of different linguistic features. A recent meta-analysis of 20 primary peer-reviewed studies has revealed a credible but small effect of predictability on pronominalization in sentence-completion tasks (Liao et al., 2025). The authors suggest that the mixed results that have been reported in these studies could be explained by the small magnitude of the effect, making it more plausible that not all studies would be able to detect it.

2.2. Predictability of events as a form of world knowledge

So far, we have reviewed studies on predictability as understood in the language sciences. However, independent evidence from cognitive science suggests that predictability is fundamental to human cognition taken more broadly, independent of language (Baldwin and Kosie, 2021; Butz, 2017; Butz et al., 2021; Kuperberg, 2021; Zacks et al., 2007). Expectations about events, their course and structure, form one of the central components of world knowledge (Butz et al., 2021). Several lines of work in psycholinguistics have addressed links between event cognition and language, investigating how speakers perceive descriptions of events involving motion (Lee and Papafragou, 2024; Papafragou et al., 2008), as well as how event roles and event boundaries are encoded in language (Ünal et al., 2021).

In descriptions of events, event predictability is often confounded with the predictability of words driven by linguistic constraints, such as the structural position of the potential antecedent, its thematic role and grammatical function, its topicality, as well as patterns of word co-occurrence. In computer science, the challenge of eliminating the confounding effect of linguistic predictability has led to the creation of special types of sentences, known as Winograd schemas.

- (3) The city councilmen refused the demonstrators a permit because they ...
 - a. feared violence.
 - b. advocated violence.

Winograd schemas feature minimal pairs of sentences where pronoun disambigua-

tion depends on world knowledge rather than tracking word statistics (Levesque et al., 2012; Winograd, 1972). Listeners are able to disambiguate the pronoun *they* as referring to the councilmen in (3a) and to demonstrators in (3b). According to Hobbs et al. (1993), pronoun resolution is driven by the search for discourse coherence, i.e., consistency with world knowledge. Such coherence relations have been argued to play a pivotal role in pronoun disambiguation (Elman et al., 2006; Kehler, 2002; Stevenson et al., 1994).

More generally, work on the role of world knowledge in pronoun disambiguation has shifted the focus of theoretical analysis from purely linguistic predictability to information predictability. For instance, constraints that were previously viewed as lexical have been grounded in world knowledge. Accordingly, Pickering and Majid (2007) argued that implicit causality verbs reflect abstractions that we have over the type of reasons that are likely to be named for particular events. Linguistic and information predictability represent distinct levels of analysis, even though linguistic predictability, at least partially, may reflect information predictability, as the example of implicit causality verbs shows (Guan and Arnold, 2021). Both analytic levels affect pronoun resolution (Kehler and Rohde, 2013). In this paper, we focus on *event predictability* as a specific type of information predictability. In our experiments, we reduce the confounds between event predictability and linguistic predictability by measuring and manipulating participants’ expectations about event transitions in non-linguistic tasks.

2.3. *Rational communication under uncertainty*

If event predictability indeed affects the disambiguation of anaphoric references, we would expect rational cooperative speakers to prefer unambiguous forms of reference when describing surprising events. In pragmatics, this link between speakers’ choices and listeners’ behavior is captured within the Rational Speech Act (RSA) model of communication (Frank and Goodman, 2012; Franke and Jäger, 2016; Goodman and Frank, 2016). This model views speaking as a process of choosing linguistic forms that maximize the chance that the listener will recover the intended message. In other words, Gricean models emphasize that understanding speech production necessarily involves modeling speech perception. Here, the effect of world knowledge on speech perception can be seen as part of a Bayesian process of combining evidence with prior expectations (Achimova et al., 2024; Kehler and Rohde, 2013; Patterson et al., 2022). Choosing optimal linguistic forms in the context of anaphoric references involves taking into account the ambiguity of such forms and their likelihood of being misinterpreted—disambiguated towards an unintended antecedent.

One possible source of such misinterpretation is the plausibility of the unintended interpretation. In the process of disambiguation, listeners combine bottom-up information with top-down prior expectations to form an interpretation, weigh-

ing priors and signals depending on their strength and reliability, respectively (Gibson et al., 2013). Reliance on priors is generally a solution that makes speech comprehension more robust: priors allow listeners to recover meaning from an unreliable signal. Evidence from computational models of the lexicon further suggests that agents with strong shared priors over the meaning space are able to rely on smaller lexicons, where ambiguous forms can be easily disambiguated by context (Brochhagen, 2020). However, if the intended meaning is overall implausible, priors may bias the listener towards an unintended interpretation, particularly if the signal is not fully reliable. The interplay between priors and linguistic signal is addressed by the noisy-channel model of communication, which offers a computational solution to dealing with a partially unreliable signal to derive meaning (Benz, 2012; Bergen and Goodman, 2015; Levy, 2008). Recent experimental evidence suggests that listeners flexibly combine prior expectations with a noisy signal, profiting from high predictability and relying on sentence context more when the signal is noisy (Van Os et al., 2022).

One form of noise that impacts communication is actual auditory interference, such as background sounds or overlapping speech from others. In a noisy environment, listeners rely heavily on their prior expectations, such as the plausibility of certain situations or utterances, to interpret ambiguous input and reconstruct the intended message (Boothroyd and Nittrouer, 1988; Hannemann et al., 2007; Miller et al., 1951; Obleser and Kotz, 2010; Sohoglu et al., 2012). Studying the interpretation of referring expressions in noisy conditions creates an environment in which we can observe the effect of prior expectations on the disambiguation of anaphoric forms.

2.4. Artificial-world paradigm

In previous work, we presented preliminary evidence that speakers take event predictability into account when referring to event participants (Achimova et al., 2022; Stegemann-Philipps et al., 2021). There, we introduced an artificial-world paradigm, in which participants first learned about possible event progressions in this world during training, and then described events with imaginary characters presented on screen. When one character attacked another and then one of the characters fled, speakers were more likely to use a definite NP to refer to the fleeing character when the outcome was surprising than when it was familiar. However, the effect was modest: when the antecedent was in the subject position, participants produced definite NPs on 62% of familiar event trials versus 70% of surprising event trials; when the antecedent was in the object position, they produced definite NPs on 96% and 98% of trials, respectively. The modest effect may partly reflect limitations of the earlier experimental design. Because familiar and surprising trials were fully randomized, surprising trials could occur early, before participants had consolidated the event expectations established during training.

Dynamic adaptation of event expectations over the course of the experiment may have reduced the distinction between later familiar and surprising trials. In addition, these studies examined only one type of interaction (attack), making it unclear whether the event predictability effect generalizes across event types.

Despite these limitations, Stegemann-Philipps et al. (2021) established a paradigm that allows experimentally manipulating event predictability and tracing its effects on speakers’ choices of referring expressions. In standard sentence-completion tasks, speakers must decide *which* referent to mention and *how* to refer to it. In the artificial-world paradigm, the visual event determines the referent to be described, isolating the “how” question, that is, the choice of referring expression. Subsequent work from our group has also offered preliminary evidence that speakers favor less ambiguous forms, such as definite NPs, over potentially ambiguous reduced forms when describing implausible events, and that, under auditory noise, listeners may sometimes interpret the degraded signal in favor of a more plausible event interpretation (Achimova et al., 2024). The experiments in the present article build on this paradigm to evaluate the effect of predictability with novel characters across a larger set of possible interactions and implement a tighter control over possible experimental confounds.

2.5. Contribution

This paper investigates whether and how event predictability shapes the production and interpretation of anaphoric references. We test two complementary predictions. First, speakers should choose more phonologically overt referring expressions when describing surprising events, because a less overt expression may be reinterpreted in favor of a more plausible event continuation. Second, we predict that listeners will rely more strongly on prior expectations about event outcomes as the linguistic signal becomes less reliable. We ground these predictions in two complementary frameworks. RSA models of communication would predict that speakers who are sensitive to possible listener interpretations should favor more overt forms when describing surprising events, precisely because surprising events raise the risk of misinterpretation: when the intended event outcome is less plausible, listeners may resolve an ambiguous or reduced form toward an unintended antecedent that supports a more plausible interpretation. The noisy-channel model predicts, complementarily, that listeners combine the linguistic signal with prior expectations about plausible event outcomes, and that this reliance on priors increases as the signal degrades, sometimes yielding an interpretation inconsistent with the presented linguistic form.

To evaluate these predictions, we expand on our prior work by sampling a wider range of interactions and by treating event predictability, with respect to the different interactions, as a continuous rather than categorical predictor. We establish

prior non-linguistic expectations over event outcomes in a norming study. In Experiments 1a and 1b, we examine speakers’ choice of zero anaphors, pronouns, and definite NPs as a function of event predictability. In Experiments 2a and 2b, we investigate how listeners interpret referring expressions as a function of event plausibility, linguistic form, and auditory noise. Across production and comprehension, we test whether speakers choose more overt forms of reference when they describe more surprising events, and to what extent listeners rely on event plausibility when interpreting speech presented with and without added auditory noise.

3. Data availability

Anonymized data and analysis code required to reproduce the results reported in this paper are available via the link: https://osf.io/nr9py/files?view_only=0db29dd5eeeb4f6f943ae22292418e4f.

4. Experimental paradigm and norming experiment

In this paper, we investigate how expectations about the outcome of an event interact with linguistic constraints on pronominalization in production and comprehension of referring expressions, including comprehension under auditory noise. To control the prior expectations of participants regarding the outcomes of different interaction events, we built on the artificial-world paradigm introduced in Stegemann-Philipp et al. (2021).¹ In our study, the artificial world depicted short cartoons featuring three kinds of monsters: red, yellow, and blue (Figure 1). The monsters engaged in one of four types of interactions (attacking, rock throwing, waving, and jumping over)² that were followed by a second, ‘outcome’ event (one monster falling down). In a norming study, we first obtained the event outcome priors for different actions. We then investigated how those priors affect the production (Experiments 1a and 1b) and interpretation (Experiments 2a and 2b) of referring expressions.

Participants in our experiments received compensation for their time at a rate of 10 British pounds per hour (£1-2 per experiment, as each experiment lasted between 5 and 10 minutes). The experiments were administered via a JATOS server (Lange et al., 2015). The experimental protocols have been reviewed by the

¹The original study used cells and viruses as the interacting entities. During their interaction, one of the characters attacked the other, which was followed by one of the entities fleeing.

²The practical considerations of having a manageable set of trials and a reasonable number of participants made us restrict ourselves to a limited number of interactions. In choosing interactions, we tried to sample those that show different biases in event outcomes. In the future, we hope to expand the range of interaction types to further increase the generalizability of our findings.

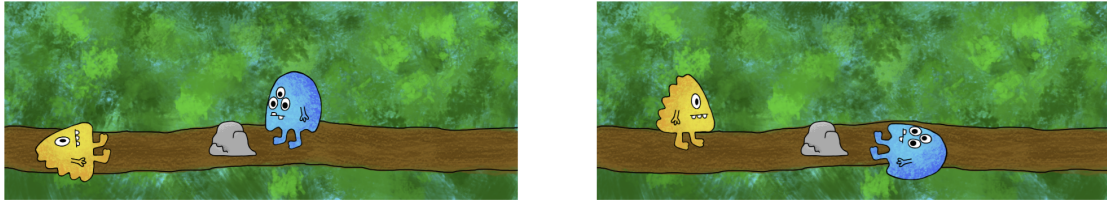


Figure 1: In our artificial world, the videos show two monsters engaged in a two-event interaction sequence. For example, the yellow monster throws a rock at the blue monster (first event) and then either the yellow monster itself (the agent of the action, left) or the blue monster (the patient of the action, right) falls to the ground (second event, or outcome).

Ethics Committee for Psychological Research at the University of Tübingen, and all participants provided informed consent.

4.1. Methods

In the norming experiment,³ our goal was to establish baseline measures of participants’ prior expectations about event outcomes for different actions. During the experiment, participants watched two scenes featuring the same pair of monsters engaged in the same type of interaction, differing only in the outcome: in one scene, the agent of the first event (throwing a rock in Figure 1) fell to the ground (left panel), whereas in the other, the patient did (right panel). Participants had to use a continuous slider to indicate which event outcome they found more plausible and to what extent (Figure 2), where the central position of the slider indicated that the outcomes were equally plausible. Each participant evaluated all four actions in random order, one trial per action, with the colors of the monsters randomly selected in each trial. Participants received no linguistic input apart from instructions. This allowed us to isolate participants’ world knowledge from their linguistic knowledge of word co-occurrences.

We recruited 50 participants on Prolific. Data from participants who indicated that they may not have understood the instructions ($n = 2$) were excluded. Data from 48 participants were included in the analysis.

4.2. Results

Figure 3 shows the mean values of how likely it is for the monster affected by the first action, that is, the patient, to fall. We treat these values as a measure of the patient bias, understood broadly as a bias that the affected monster would fall, and use these values as a continuous predictor in subsequent experiments. A value

³The experiment is available via the link:
<https://www.cognitive-modeling-experiments.uni-tuebingen.de/publix/insW0r74blh>

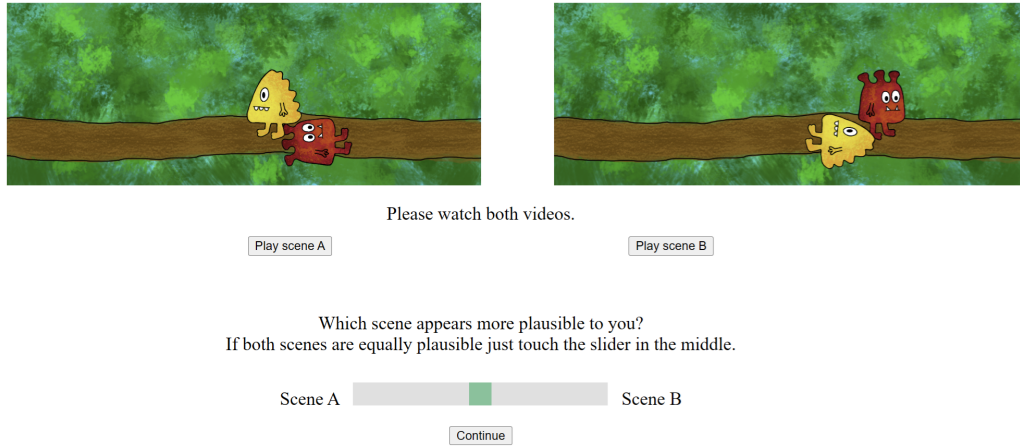


Figure 2: A sample trial of the norming study. In the example, the red monster just jumped over the yellow monster (interaction event), followed by the red monster (left) or the yellow monster (right) falling over (outcome).

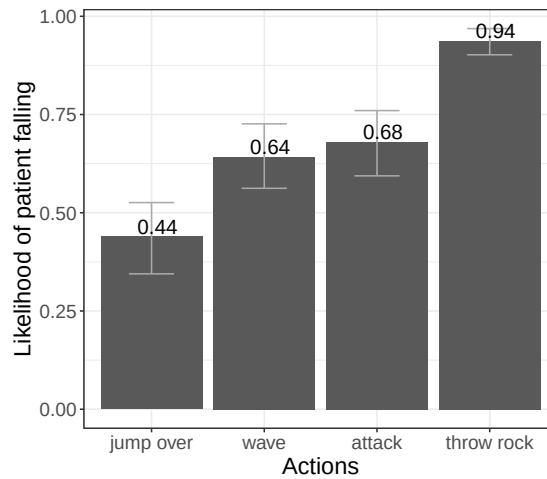


Figure 3: Norming study. We use the mean values of patient bias as action priors. These values reflect how likely it is that the patient rather than the agent of the first event will fall as a result of the interaction. Error bars here and in subsequent plots correspond to 95% confidence intervals.

of 1 means that the monster affected by the first interaction (e.g., the patient of *attack*) is likely to fall down as a result, while a value of 0 means that the agent of the first event is likely to fall. The results showed that the *throw rock* action yields the strongest patient bias, the *attack* and *wave* actions have intermediate biases, and *jump over* action has a weak agent bias or no bias at all. The four types of interactions thus cover a meaningful continuum of possible interaction outcome

predictions, which we refer to as *action priors*.

5. Experiments 1a and 1b: Speech production

Experiments 1a and 1b focus on the question to which extent the predictability of event outcomes influences utterance choice. More specifically, we ask which type of a referring expression (a zero anaphor, a pronoun, or a definite NP) participants will produce when referring to the falling character. To evaluate the effects of event predictability and structural constraints (position of the antecedent) on the choice of referring expression for actions that differ in their natural action priors over possible outcomes, we used four actions introduced in the norming study (*jump over*, *wave*, *attack*, and *throw rock*). Additionally, we induced experimental priors by familiarizing participants with typical interaction outcomes during a training period. Experiment 1b examines if the training was indeed responsible for the observed results by dropping the familiarization part.

5.1. Experiment 1a: Speech production with trained event-level priors

In Experiment 1a,⁴ we investigated how speakers choose referring expressions to refer to characters anaphorically depending on whether the event outcome is familiar or surprising. We used the action events assessed in the norming study that naturally differ in the strength of the expectation that the patient of the event should fall down as a result. In order to investigate the potential interaction between event predictability and antecedent position, we additionally formed experiment-specific expectations by associating the likely outcome with monster colors that marked their strength. This manipulation allowed us to create a condition that made the *patient-falling* outcome more likely for particular color combinations of monsters, and another condition where the *agent-falling* outcome was more plausible. In the production phase, we exploited these learned event priors and created animations of familiar and surprising event outcomes—those that conformed to or violated the expectations over outcomes that we previously formed. Effectively, this manipulation created an environment where we can assess the impact of event-level predictability (including action priors and learned outcome expectations) and antecedent position (agent or patient) on the type of referring expressions that speakers produce.

⁴The experiment was preregistered on OSF: https://osf.io/dhz8f?view_only=d1bb6a9bf58a42f8a8d851e2b4e6b540 and is available via the link: <https://www.cognitive-modeling-experiments.uni-tuebingen.de/publix/zk6MTkF1hej>

5.1.1. Methods

Participants were randomly assigned to one of the four actions: jump over, wave, attack, or throw a rock (between-subjects design). The experiment included three phases: training, a memory test, and a free-production task. In the training phase, participants learned to recognize the relative monster strength by observing several monster interactions and their outcomes. To induce expectations over event outcomes, we fixed the color of the monster who initiated the interaction (yellow) and manipulated the color of its partner, such that the partner’s color predicted who would fall down. In particular, the yellow monster was stronger than the blue one, causing the blue one (the patient) to fall, while the red monster was stronger than the yellow one, causing the yellow one (the agent) to fall. After the training phase, we evaluated participants’ knowledge of the monster strength. Participants watched a scene featuring two monsters at the start of their interaction and picked one of the two possible outcomes as the more likely one (Figure 4); participants completed two of such trials. If they failed at least one of the tests, they were asked to repeat the training phase. If they failed the knowledge test the second time, they were still allowed to complete the experiment and receive reimbursement, but their data were excluded from the analysis.

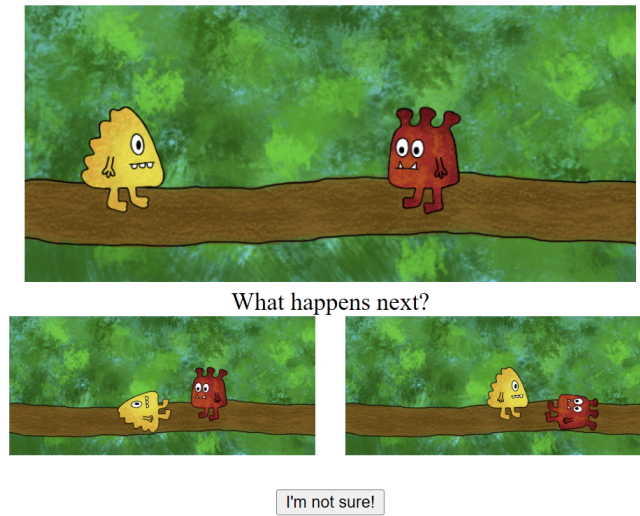


Figure 4: Experiment 1a. A sample knowledge test trial: A yellow monster attacks a red monster. As a result, either the yellow (left) or the red (right) monster falls down. The participant had to select one of the two possible outcomes of the interaction or click the “I’m not sure” button.

In the main free-production phase, participants observed an interaction between two monsters and were asked to describe what they saw, speaking into a microphone. Participants were instructed to simply describe what was going on and to speak in a normal, casual way as they would to a friend. Each participant

completed eight trials: six critical trials and two fillers. The first and third trials were fillers featuring monsters without any interaction. The remaining six trials were structured as follows: the first two involved familiar events, where the outcome conformed to the relative strength of the monsters learned during the training phase. The next two trials presented surprising events, where the outcome contradicted the learned strength of the monsters. The final two trials again conformed to the trained strength of monsters. We termed those trials *post-surprising*, since they followed the surprising block. At this point in the experiment, expectations over event outcomes were likely blurred, since participants experienced that in this artificial world, events did not always conform to the expectations they had previously formed during training, making event outcomes less predictable. Each block contained one agent-outcome trial and one patient-outcome trial, with their order randomized within the block. The structure of the experiment is presented in Figure 5.

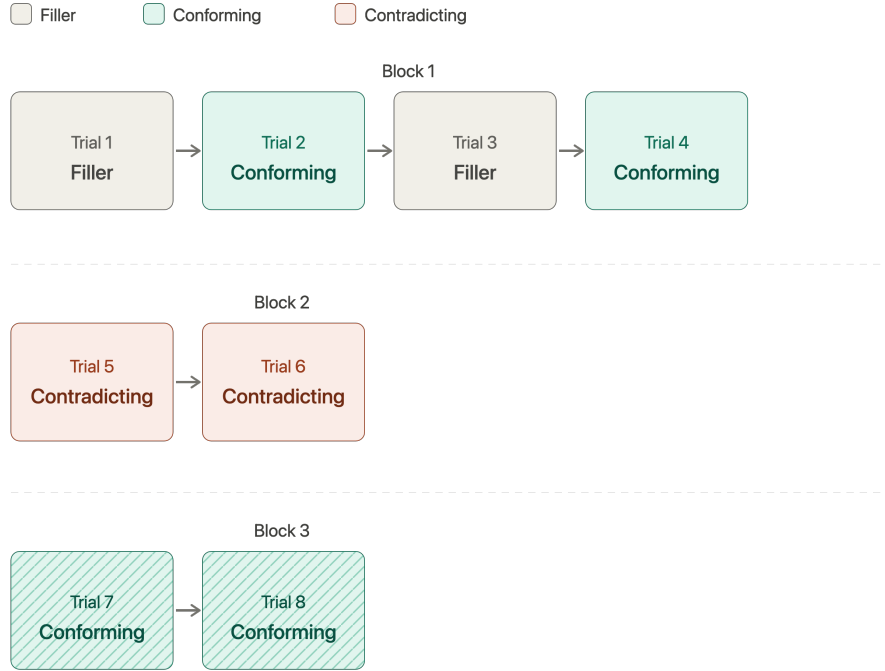


Figure 5: The structure of the main phase of the production experiment showing whether the outcome of each trial conforms to the learned monster strength or contradicts it. The diagonal hatching in Block 3 indicates the blurred familiarity of these events compared to Block 1.

Participants’ speech was transcribed and annotated according to the type of referring expression used: definite NP, pronoun, or zero anaphor. Only utterances that matched a specific pattern were considered: “The yellow monster attacked

the red monster and [(zero)/it/the yellow monster] fell over.”

Hypotheses. If speakers optimize their choice of anaphoric references to ensure that listeners can recover the intended referent, participants should use more overt, unambiguous expressions, such as definite NPs, to refer to characters of surprising events (Block 2), while, for familiar events (Block 1), they should opt for shorter forms, such as pronouns or zero anaphors. This effect is expected to be more pronounced for references to agent antecedents, as previous studies have shown that speakers primarily use noun phrases to refer to the antecedent previously mentioned in the object position (Bott and Solstad, 2023; Brennan, 1995; Demberg et al., 2023; Rohde and Kehler, 2014). This prediction regarding the effect of predictability for familiar compared to surprising events is part of our preregistration.

Beyond the preregistered predictions, the current design also allows us to address two further exploratory questions about the effect of different types of event priors and their dynamic update. In addition to measuring the effect of event predictability that we induced within the experiment, we also tracked how the priors measured in the norming study influence speakers’ choices. Specifically, stronger expectations that the *patient* of the first action would fall should lead speakers to more frequently use definite NPs when they describe a scene where the *agent* falls instead. Furthermore, the comparison of conforming trials in Block 1 and Block 3 allows us to investigate whether speakers adjust their priors in the course of the experiment. If priors shift dynamically, we expect participants to produce more overt referring expressions for post-surprising trials (Block 3) compared to familiar trials in the beginning of the experiment (Block 1). This effect would explain why in earlier production experiments, where the order of trials was not controlled (Stegemann-Philipps et al., 2021), the effect of predictability was small.

Pilot experiment and power estimation. To calculate the number of participants necessary to assess the effect of predictability on the choice of referring expressions, we used estimates of this effect previously obtained in a pilot study presented in Achimova et al. (2024). In the pilot study ($n = 44$), we investigated whether event predictability affects the probability of using a definite NP compared to a reduced form of reference. The experiment followed the same experimental paradigm, as described in the current work (six critical trials in three blocks: familiar, surprising and post-surprising), but was restricted to three interactions: *attack*, *throw a rock*, and *jump over*. The analysis in our pilot study focused on the contrast between familiar and surprising events and excluded post-surprising trials from the calculation of the effect of predictability. To analyze this effect, we regressed the type of referring expression (definite NP vs. reduced form) on the type of event (familiar vs. surprising) separately for agent and patient antecedents, focusing on the trials where the anaphor antecedent was the agent of

the first event, since for patient antecedents participants overwhelmingly produced definite NPs. For agent antecedents, we registered the effect of event predictability: participants produced more definite NPs for surprising compared to familiar events ($\beta = 18.978, SE = 3.771, z = 5.033, p < 0.01$). We use this coefficient for sample size estimation for the current study reflected in our preregistration. We further calculated the regression coefficient for the patient-antecedent condition. The analysis reported in the pilot study did not take into account the interaction between event predictability and the position of antecedent, as well as the contribution of action priors.

We performed simulation-based power analysis to estimate the effect of event predictability (familiar vs. surprising, treatment-coded) on the type of referring expression produced (reduced form vs. definite NP, coded as a binomial variable).⁵ Power was estimated using the *simr* package (Green and MacLeod, 2016) in R (R Core Team, 2021), with the design matrix adjusted to add one further action type. We then fit binomial mixed-effects models predicting the choice of referring expression (NP vs. reduced form) as a function of event predictability, estimating the corresponding regression coefficients from only the first four target trials and excluding post-surprising events. Following Achimova et al. (2024), we ran the power simulation separately for patient and agent antecedents; it did not cover the post-surprising comparison, action-prior effects, interactions involving antecedent position, or the exploratory three-level ordinal analysis reported below.

The results indicated that a sample size of 200 participants would provide 80% power to detect the effect of event predictability on the form of referring expression for patient antecedents, assuming the effect size observed in the original data. For agent antecedents, the analysis showed that a sample of 200 participants would yield close to 100% power, even when the effect size was reduced by half.

Participants. Based on the power estimation, we recruited 200 participants on Prolific. 24 participants were excluded (1 participant specified that their native language is not English, 6 indicated that they might not have understood the instructions and 17 failed the memory test twice). Additionally, trials with missing audio or with utterances that did not match the specified pattern were filtered out. This led to the exclusion of 37 more participants who had no usable trials. A total of 699 utterances from 139 participants were included in the analysis. The number of analyzed trials per action, trial block, and antecedent position ranged from 15 to 37, with a mean of 29.125.

⁵Model syntax: `glmer(expression ~ event type + (1 | subject id))`.

5.1.2. Model estimation

To determine the maximal random-effects structure justified by the data, we first conducted model comparisons using a likelihood ratio test, varying the random-effects structure of the maximal model that included interactions between all fixed predictors. Once the random-effects structure had been identified, we performed a second model comparison using a likelihood ratio test to detect the maximal fixed-effects structure supported by the data. Including random slopes for all the predictors by participants often led to a model in which the number of random effects exceeded the number of observations. This issue is referred to as over-parameterization and is a common problem in mixed-effects modeling when the number of items is relatively small (Bates et al., 2015). The same model selection procedure was applied in Experiment 2a, while the models in Experiments 1b and 2b were selected to mirror those in Experiments 1a and 2a, respectively.

5.1.3. Results

Based on reviewers’ recommendations, we adjusted our analysis plan and performed a combined analysis on the full dataset, including both agent and patient antecedents. This allowed us to estimate the well-documented effect of antecedent position, i.e., subjecthood (Arnold, 2001, 2010; Brennan, 1995; Demberg et al., 2023; Grosz et al., 1995) as well as its potential interaction with event type and action priors. We further chose to treat the type of referring expression produced as an ordinal three-level variable to minimize data post-processing. Deviations from the preregistered protocol are documented in Appendix A.1.

The data were fitted with a cumulative link mixed model implemented in the *ordinal* package (Christensen, 2026). We treated the produced referring expression as a categorical dependent variable, ordered by the level of phonological overtiness (zero anaphor < pronoun < definite NP). Action priors, event type, and antecedent position served as fixed effects; we further assessed the only interaction between action priors and antecedent position, as adding other interactions did not improve the model fit. The random-effects structure contained random intercepts per participant and item⁶ as well as a random slope for antecedent position per participant, as justified by the model comparison. Familiar events and agent antecedents served as reference levels.

The analysis revealed a significant simple effect of action priors ($\beta = 5.00$, $SE = 1.19$, $z = 4.19$, $p < 0.001$), with higher action priors associated with more overt forms of reference for agent antecedents (Table 1, Figure 6). Antecedent position likewise reached significance ($\beta = 11.13$, $SE = 2.73$, $z = 4.08$, $p < 0.001$),

⁶An item here and in subsequent experiments corresponds to the colors of interacting monsters.

	β	SE	z	p
event (surprising)	1.48	0.31	4.72	< 0.001
event (post-surprising)	1.23	0.30	4.04	< 0.001
action prior	5.00	1.19	4.19	< 0.001
patient	11.13	2.73	4.08	< 0.001
action priors : patient	-6.75	3.36	-2.01	0.045

Table 1: Experiment 1a (full dataset). Produced referring expression (zero anaphor, pronoun, or NP). Model syntax: `clmm(expression ~ event type + action prior * antecedent + (1 + antecedent | subject id) + (1 | item))`. Familiar events and agent antecedents serve as reference levels.

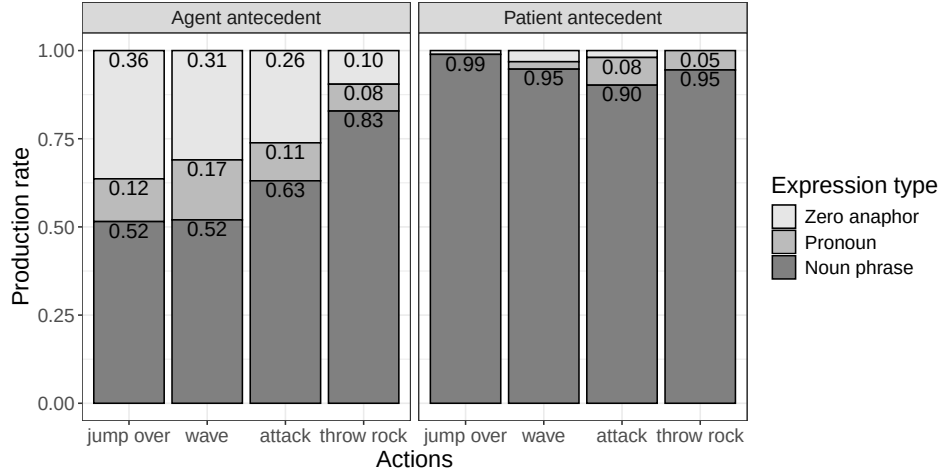


Figure 6: Experiment 1a (full dataset). Produced referring expressions by action and antecedent position.

indicating that patient antecedents elicited more overt referring expressions than agent antecedents. Finally, the interaction between action priors and antecedent position was significant ($\beta = -6.75$, $SE = 3.36$, $z = -2.01$, $p = 0.045$), reflecting the opposite and attenuated effect of action priors for patient antecedents relative to agent antecedents (see also Table A.8, where the reference level was changed to patient antecedents to test the simple effect of priors in this case).

The ordinal regression model reported above demonstrated that the likelihood of more overt referring expressions increased for actions with stronger patient bias. In order to evaluate which of the contrasts were responsible for these effects, we conducted a post-hoc analysis with the Tukey method for multiple comparisons, evaluating for each of the referring expressions whether the rates differ between different actions. The inventory of referring expressions differs for agent antecedents—they allow all three forms of anaphoric reference, while patient antecedents grammatically only allow two forms: a pronoun, and a definite NP. It

means that for patient antecedents the main ordinal model’s contrasts are already equivalent to a binary comparison; the post-hoc decomposition was therefore only necessary for the three-category agent-antecedent condition. We estimated model-predicted marginal probabilities of producing each form of anaphoric references with the *emmeans* package (Lenth and Piaskowski, 2026) for the agent-antecedent condition. The post-hoc analysis revealed that the probability to produce a definite NP was higher and the probability to produce a pronoun or a zero anaphor was lower for actions with stronger patient bias (Table A.9).

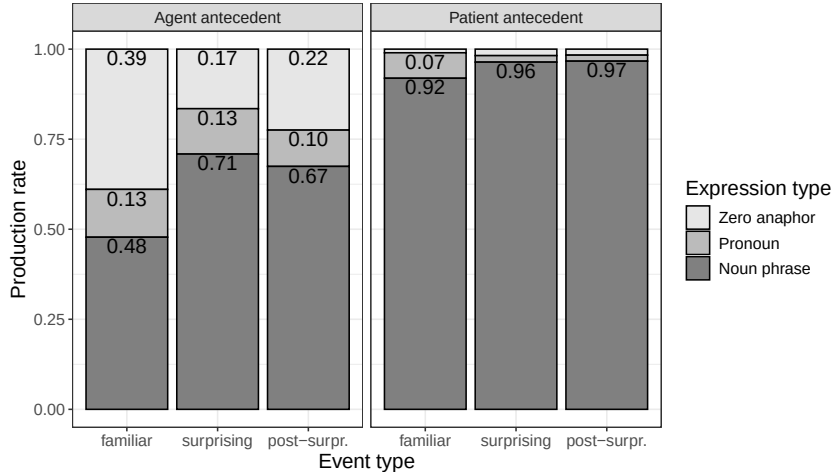


Figure 7: Experiment 1a (full dataset). Produced referring expressions by event type and antecedent position.

In line with our preregistered hypothesis, event type significantly affected the overtiness of referring expressions: participants were more likely to produce more overt forms both for surprising events (testing our preregistered hypothesis within this exploratory unified model; $\beta = 1.48$, $SE = 0.31$, $z = 4.73$, $p < 0.001$) and for post-surprising events (exploratory hypothesis; $\beta = 1.23$, $SE = 0.30$, $z = 4.04$, $p < 0.001$) relative to familiar events (Table 1, Figure 7). In addition, the model with surprising events (instead of familiar) as the reference level (Table A.7) showed that post-surprising events were not significantly different from the surprising ones.

5.1.4. Discussion

Speakers produced more overt expressions when referring to the patient of the first event than when referring to the agent, reflecting a well-documented effect of subjecthood (Arnold, 2001, 2010; Brennan, 1995; Demberg et al., 2023; Grosz et al., 1995). We attribute this pattern to differences in referent salience (Arnold, 2010). After the first event, the agent is likely the most salient referent. Besides its visual salience (as the moving entity), its prominence is also supported by several

linguistic factors known to increase the rate of pronominalization: the agent was consistently the subject and the first-mentioned entity (Arnold, 2001; Brennan, 1995), and it was likely the discourse topic (Kehler and Rohde, 2013). Our findings are therefore consistent with the previous studies: participants were more likely to produce a pronominal reference for subject rather than object antecedents. However, given our experimental design, we are unable to fully disentangle the effects of topichood, thematic role, and grammatical position. These factors may exert distinct pressures on pronominalization (Arnold, 2001; Hwang et al., 2022; Rosa and Arnold, 2017).

In this experiment, our focus was on evaluating the role of both natural action priors and experiment-induced monster strength priors. We observed that for actions with stronger patient bias, participants were more likely to use more overt forms of reference to refer to the agent (Figure 6, left panel). The training phase did not replace the participants’ natural action priors (their world knowledge), with both types of priors carrying their own independent effects. We conclude that while event semantics, as revealed in action priors, captures general expectations over particular event outcomes, these expectations can be further affected by situation-specific knowledge. In our experimental set up, for example, the likelihood that one of the event actors will undergo a second event was affected by their relative strength. We have demonstrated that both event semantics (action priors) and situation-specific knowledge (monster strength) are relevant for the choice of referring expressions. More specifically, speakers produced more overt referring expressions for event outcomes that were more surprising either from the point of view of action-specific priors or the expectations over monster strength developed within the experiment.

We also notice that while the use of noun phrases increased in surprising events and more patient-biased actions, the use of zero anaphors for agent antecedents decreased (Figures 6 and 7, left panels). This effect is compatible with the idea of speakers reasoning about their potential listeners when choosing utterances. First, speakers anticipate that listeners may interpret ambiguous pronouns according to their prior expectations. When the outcome contradicts these expectations, speakers tend to use more overt forms, such as definite NPs, to prevent the default interpretation. Second, although zero anaphors are inherently unambiguous, speakers seem to anticipate that in noisy conditions listeners might misinterpret them as ambiguous pronouns, leading them again to rely on their priors for interpretation. In Experiments 2a and 2b we will test this hypothesis from the listener side.

Our exploratory analysis of post-surprising events suggests that priors over event outcomes shifted dynamically in the course of the experiment: as participants encountered more unexpected outcomes, their priors likely changed, making event outcomes less predictable overall. Although these post-surprising trials featured

the same outcomes as the first two familiar trials, participants’ referring expression patterns aligned with those used for surprising rather than familiar events.

We have so far attributed the difference between surprising and post-surprising events compared to familiar ones to the effect of event predictability and the dynamic event-specific prior update, respectively. However, since surprising and post-surprising trials always occurred later in the experiment, we cannot exclude the possibility that this was simply the effect of trial order. We therefore conducted a follow-up experiment where we examined whether the distinctiveness of familiar trials, which preceded surprising and post-surprising trials in the experiment, was attributable to training rather than trial order alone.

5.2. Experiment 1b: Speech production without prior training

To test whether the difference between familiar and surprising events in Experiment 1a depended on the expectations established during training, we conducted a follow-up experiment without the training and memory-evaluation phases.⁷ The analysis reported for this experiment is exploratory.

5.2.1. Methods

The design of this experiment mirrored that of Experiment 1a. However, this time participants skipped the training and memory phases and proceeded directly to the main production phase. The trials were arranged in the same order as in the main experiment (Figure 5), however this time the fictive monster strength was unknown to the participants.

Hypotheses. If the effect of event type observed in Experiment 1a was due to trial order alone, we should see the same effect in the experiment without training since it features the same order of trials. But if we observe no effect or a reduced effect size in Experiment 1b, that will indicate that training significantly affected participants’ choices in Experiment 1a.

Participants. The number of conditions was identical to that of the main experiment, so we recruited 200 participants on Prolific. 11 participants were excluded (3 specified that their native language is not English and 8 indicated that they might not have understood instructions). Additionally, trials with missing audio or with utterances that did not match the specified pattern were removed. This led to the exclusion of 25 more participants who had no usable trials. In total, 774 utterances from 164 participants were included in the analysis. The number of analyzed trials per action, trial block, and antecedent position ranged from 18 to 39, with a mean of 32.25.

⁷The experiment is available via the link:
<https://www.cognitive-modeling-experiments.uni-tuebingen.de/publix/BA3tm3SE6es>.

5.2.2. Results

The results of Experiment 1b are presented in Figure 8 (compare to Figure 7). For agent antecedents, participants produced 34% of definite NPs in the first block of trials, compared to 48% for the second block, and 54% for the third block. This pattern suggests that trial order may contribute to an increase of more overt, less ambiguous descriptions, such as definite NPs, in later blocks of trials.

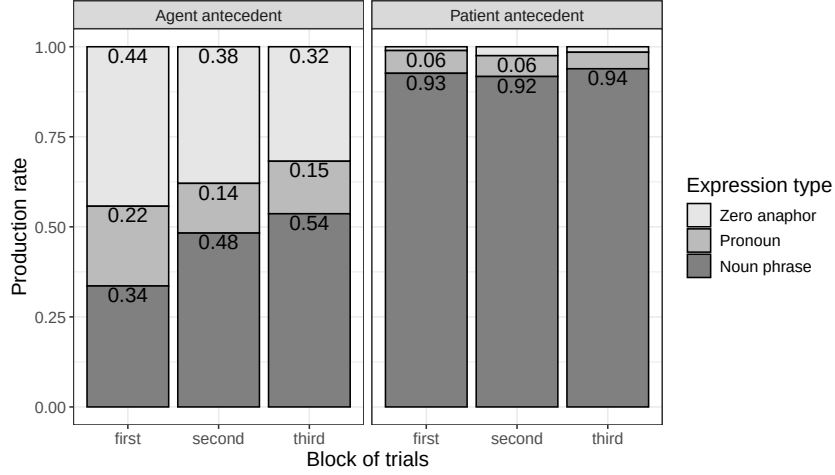


Figure 8: Experiment 1b (full dataset). Produced referring expressions by block of trials and antecedent position.

	β	SE	z	p
trial block (second)	0.57	0.26	2.15	0.031
trial block (third)	0.97	0.26	3.71	< 0.001
training	0.59	0.33	1.80	0.072
action prior	5.93	0.80	7.46	< 0.001
patient	11.06	1.62	6.83	< 0.001
trial block (second) : training	0.93	0.39	2.37	0.018
trial block (third) : training	0.28	0.39	0.73	0.468
action priors : patient	-6.55	2.18	-3.00	0.003

Table 2: Experiments 1a and 1b (combined dataset): produced referring expression (zero anaphor, pronoun, or NP). Model syntax: `clmm(expression ~ trial block * training + action prior * antecedent + (1 + antecedent | subject id) + (1 | item))`. First block of trials, agent antecedents and the absence of training serve as reference levels.

We analyzed the two experiments jointly in a single combined model. With the absence of training and the first block of trials as reference levels, the trial-block coefficients estimate the within-Experiment-1b block contrasts, whereas the

block-by-training interactions test whether these contrasts differed between Experiments 1a and 1b. We fitted a cumulative link mixed-effects regression model testing for the effect of the interaction between trial block and training on the produced referring expression (coded as a three-level ordinal variable: zero anaphor < pronoun < definite NP). The model also included an interaction between action priors and antecedent position, random intercepts for participants and items, and random slopes for antecedent position by participants (mirroring the model used in Experiment 1a).

The analysis (Table 2, first two rows) revealed a simple effect of the second and third blocks of trials relative to the first one in the experiment without training, as well as a significant interaction between the second block and the presence of training (Table 2, sixth row). In other words, even though changes across trial blocks occurred without training, the contrast between the first and second blocks of trials was significantly larger in Experiment 1a, indicating an effect of training. However, the interaction between the third block of trials and the presence of training was not significant (Table 2, seventh row).

5.2.3. Discussion

Despite the absence of training in Experiment 1b, we observed a difference in referring expression production between the first and the second blocks of trials, as well as between the first and the third blocks. Since participants were not trained to recognize which events were familiar, this pattern indicates that block changes can occur without training, reflecting a general adaptation to the observed event outcomes: as participants realize that the same interaction can lead to two different outcomes and that the outcomes are difficult to predict, they prefer more overt and less ambiguous expressions to refer to the actors of those events.

Nonetheless, the increase in referential overtiness from the first to the second block was significantly larger in Experiment 1a than in Experiment 1b. This interaction provides evidence that the training intervention causally contributed to the production pattern: experimentally induced expectations about monster strength amplified speakers' tendency to use more overt referring expressions when those expectations were violated. Because the same sequence of trials was present in Experiment 1b, but without the color-specific training, this comparison allows us to distinguish the effect of the experimental intervention from the more general effect of trial order.

The nonsignificant third-block-by-training interaction (Table 2, seventh row) offers an explanation of the mechanism behind this effect. In Experiment 1a, participants' expectations were likely reset after observing surprising events in the second block: they stopped expecting that event outcomes (which of the monsters falls down) are associated with the monster color. In Experiment 1b, throughout blocks 1 and 2 participants also experienced a variety of events where the outcomes

were not tied to monster colors. Thus, both in Experiments 1a and 1b, participants no longer had color-specific expectations over event outcomes in the third block.

6. Experiments 2a and 2b: Speech perception

In Experiments 1a and 1b, naturally occurring action priors and expectations established through training affected the participants' choice of referring expressions: speakers avoided reduced forms of reference to describe surprising events. This behavior could be seen as a strategy to reduce the risk of misinterpretation if listeners indeed interpret reduced expressions in light of their expectations about event outcomes. Experiments 2a and 2b test this comprehension prediction: we investigate (1) whether listeners use event plausibility to interpret ambiguous pronouns, and (2) whether they tend to reinterpret utterances containing zero anaphors as though a pronoun had been present if this allows them to arrive at a more plausible interpretation. In Experiment 2a, we test these predictions in an environment with auditory noise, while in the follow-up Experiment 2b, we evaluate whether the effects persist in a clean environment without a specially added auditory noise.

6.1. Experiment 2a: Speech perception under noise

The goal of Experiment 2a was to investigate how listeners combine their prior expectations with a linguistic signal while interpreting referring expressions under noisy conditions.⁸ Noisy conditions allow us, on the one hand, to simulate an ecologically plausible setting in which speech perception commonly unfolds, and on the other hand, to observe the relative weight listeners assign to prior expectations when the linguistic signal is only partially reliable.

6.1.1. Methods

As in the norming study, participants watched two scenes in which two monsters performed the same action with different outcomes (either the agent or the patient of the action fell to the ground). They then heard a statement, such as (4), in one of the two versions: the prompt either contained a conjoined verb phrase and was therefore not ambiguous with regard to the subject of the *fall down* event (4a), or it contained a pronoun *it*, as in (4b), which permitted both a reading where the agent or the patient of the previous event falls down.

(4) Auditory prompt:

⁸The experiment was preregistered on OSF: https://osf.io/5j8ua?view_only=3f6894e9f66942ae873e6e610f39238f and is available via the link: <https://www.cognitive-modeling-experiments.uni-tuebingen.de/publix/mY0tjutGJ6Q>

- a. The red monster attacked the yellow monster and fell down.
- b. The red monster attacked the yellow monster and it fell down.

We generated the prompt recordings with the default intonation contour (falling pitch) and no accent on the pronoun using a text-to-speech service ttsmp3.com. Using a single text-to-speech voice and fixed synthesis settings standardized the recordings across items. By crossing 6 different pairs of monster colors, 4 actions, and 2 prompt types (ambiguous and unambiguous), we created 48 unique stimuli. Once generated, the prompts were combined with a cocktail party noise sample adjusted for the required SNR (signal-to-noise ratio) level.

In the experiment, we manipulated the type of action, the prompt type (ambiguous vs. unambiguous), and the level of cocktail party noise (low vs. high). The low noise condition featured prompts where the noise was as loud as the signal (0 dB SNR), and the high noise condition contained prompts with noise that was 3 dB louder than the signal (-3 dB SNR). Each participant evaluated four trials, one for each action. In each trial, the monster-color pairing, prompt type, and noise level were assigned randomly.

The participants performed two tasks over the course of each trial. First, they selected the scene that best matched the description they had heard. Second, they were presented with a textual prompt that corresponded to the audio prompt that they had previously heard, but without the final part, as in (5), and were asked to type the final part based on what they had heard.

- (5) Text prompt:
The red monster attacked the yellow monster and ...

We then annotated their responses, specifying which referring expression they used: a zero anaphor, as in (4a), a pronoun (4b), or something else. Only zero anaphors and pronouns were considered in the analysis.

Hypotheses. Our predictions about the effect of priors on the choice of event outcomes differ for ambiguous and unambiguous prompts. When a pronoun is present, it can refer to the agent or patient of the previous event, making the prompts ambiguous. Thus, for ambiguous prompts, both event outcomes are compatible with the audio description. We predicted that prior expectations regarding the plausibility of event outcomes would influence the choice of the outcome. The more patient bias participants associate with a particular action, the more often they should select the scene where the patient falls as a more plausible interpretation of the audio prompt.

In contrast, the prompt without a pronoun should technically allow only the interpretation where the agent of the previous event falls. However, under noisy conditions, participants may reinterpret the signal based on their prior expect-

tations, as noisy-channel models suggest (Gibson et al., 2013; Levy, 2008), and reconstruct the pronoun if such reconstruction would allow them to relate the description to a more plausible event. For example, mentally reconstructing a pronoun *it* would turn the prompt ambiguous, permitting the pronoun to be co-referenced either with the agent or the patient of the previous event. We expected this mental revision to be more likely for actions with stronger patient bias. The second task (typing what participants heard) allowed us to test whether listeners produced responses consistent with representing the utterance as containing a pronoun more often when they had stronger expectations of the patient falling. Finally, we expected the level of noise to influence such reconstructions. The noisy-channel model predicts that under high noise, insertions of pronouns should occur more frequently, and the effect of priors should be stronger, since in those circumstances listeners will rely more on their priors to compensate for the unreliable signal.

Pilot experiment and design choices. In a pilot experiment ($n = 49$) reported in Achimova et al. (2024), we assessed whether the disambiguation of anaphoric references under noisy conditions was affected by action priors. The pilot experiment included three types of actions and prompts masked either by noise equally loud to the prompt (0 dB SNR condition) or 5 dB louder than the prompt (-5 dB SNR condition). We analyzed (1) the probability of choosing a scene where the patient of the event falls and (2) what anaphoric reference participants report as having heard, performing the analysis separately for ambiguous and unambiguous prompts. Applying treatment coding, we set *attack* as a reference level and showed that for the action with a weaker bias (*jump over*), participants were less likely to select the patient outcome ($\beta = -0.7230, SE = 0.2359, z = -3.065, p < 0.01$) and they were also less likely to reconstruct a pronoun for prompts containing no pronoun ($\beta = -0.709, SE = 0.351, z = -2.018, p < 0.05$). For *throw rock* relative to the *attack*, the probability of selecting the patient outcome was not different ($\beta = -0.2913, SE = 0.2383, z = 1.223, p = 0.22$) and participants were not more likely to add a pronoun to a prompt lacking it ($\beta = -0.266, SE = 0.327, z = 0.813, p = 0.42$). Finally, the pilot experiment revealed the effect of noise: insertions of pronouns were more likely in the high noise condition ($\beta = -1.095, SE = 0.303, z = -3.617, p < 0.001$).

The analysis reported for the pilot experiment had several important limitations. It fragmented the dataset into ambiguous and unambiguous prompts, which precluded us from analyzing the overall effect of prompt type as well as the interaction of prompt type with other factors. The treatment of action priors as a categorical predictor made it hard to generalize these pairwise comparisons of actions to other types of actions. The current experiment differed from the pilot study in the following ways:

1. The number of actions was increased from three to four to improve generalizability.
2. Rather than treating action type as a categorical predictor, we used continuous patient bias scores obtained in the norming study.
3. We reduced the number of trials per participant from 12 to 4 to minimize possible adaptation to noise within the experiment.
4. We adjusted the high noise level to -3 dB SNR in the current setup since participants reported the -5 dB SNR condition to be auditorily very challenging.

The changes introduced to the analysis plan paired with a possible change in the effect size due to a reduced number of trials and noise levels made our original sample size estimation protocol inadequate. However, we retained the preregistered target of 200 participants. For transparency, the original power estimation protocol is reported in Appendix B.1.

Participants. 200 participants completed Experiment 2a on Prolific. We filtered out data from 2 non-native speakers and 7 participants who reported confusion with the task instructions. Additionally, we excluded trials where participants typed anything other than zero anaphors or pronouns in the second task. This resulted in the exclusion of 4 more participants who had no usable trials. A total of 662 trials from 187 participants were included in the analysis. Out of 662 trials 299 featured ambiguous prompts (160 with low noise and 139 with high noise) and 363 unambiguous prompts (200 with low noise and 163 with high noise).

6.1.2. Results

We collected two types of responses: the selected outcome and the typed expression. The first response captures how listeners resolve referential ambiguity, while the second one provides an additional behavioral measure of whether participants represented the utterance as containing a pronoun.

According to our preregistered plan, we intended to capture the effect of action priors on the selected outcome for ambiguous and unambiguous prompts separately. We further aimed at evaluating the interaction effect of action priors and noise level on the typed responses and event selection for unambiguous prompts. The analysis we present here differs from the preregistration in two respects. First, we analyzed ambiguous and unambiguous prompts within a single model, instead of having separate analyses for each prompt type. Second, we used model comparison to identify the maximal random- and fixed-effects structure supported by the data. The unified model allowed us to estimate the effect of prompt type and its interactions directly, while avoiding duplication across separate models. Random- and fixed-effects selection followed the model-comparison procedure described for

Experiment 1a. The preregistered prompt-specific analyses are reported in Appendix B.2 for transparency.

Selected outcomes. Ambiguous prompts, i.e., those containing a pronoun, license both a reading where the pronoun refers to the agent of the first event and a reading where it refers to the patient. If disambiguation is affected by prior expectations, we should observe a higher selection rate for event outcomes that participants in the norming study found more plausible. For unambiguous prompts, we generally expect only the agent outcomes to be chosen, since the grammar does not license a reading where the zero anaphor is coreferential with the patient. However, if, under noise, listeners reinterpret zero anaphors and perceive them as pronouns, then they could possibly select patient outcomes for unambiguous prompts as well.

We fitted the data with a mixed-effects binomial regression model implemented in the *lme4* package (Bates et al., 2015) in R with prompt type, prior rate, and noise level as independent variables and the selected outcome (agent- or patient-falling scene) as a binomial dependent variable. The model further included random intercepts for participants and items, and an interaction term between prompt type and noise level, as justified by the model comparison.

	β	SE	z	p
(Intercept)	-5.87	0.71	-8.30	< 0.001
ambiguous prompt	2.36	0.39	6.02	< 0.001
high noise	0.80	0.40	2.00	0.045
action prior	3.95	0.71	5.53	< 0.001
ambiguous prompt : high noise	-1.06	0.51	-2.07	0.039

Table 3: Experiment 2a (full dataset). Selected outcome (agent or patient falling). Model syntax: `glmer(selected outcome ~ prompt type * noise level + action prior + (1 | subject id) + (1 | item))`. Unambiguous prompts and low noise serve as reference levels.

The analysis revealed that prior expectations over event outcomes affect disambiguation (Table 3, fourth row): participants were more likely to select the action outcome where the patient falls for actions with stronger patient bias (Figure 9). For unambiguous prompts, high noise increased the probability of selecting the patient-falling scene (Table 3, third row). The negative coefficient for the prompt-type-by-noise-level interaction (Table 3, last row) indicates that the effect of noise level was attenuated for ambiguous prompts.

The analysis suggests that under noisy conditions participants reinterpreted the unambiguous prompts and reconstructed pronouns when such reconstruction allowed them to select a more plausible outcome. Thus, the rate of the reconstruction should be modulated by plausibility priors. We seek additional behavioral evidence consistent with reconstruction in the analysis of typed expressions below.

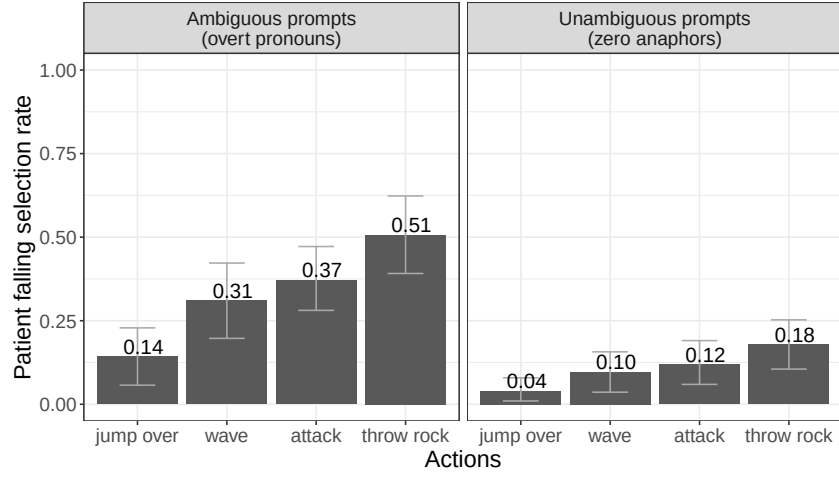


Figure 9: Experiment 2a (full dataset). Selection rate of the patient-falling scene by action and prompt type.

Expressions typed. The analysis of the second task showed a very similar pattern for pronoun typing rate as for patient outcome selection in the first task, as shown in Figure 10. Table 4 displays the results of a mixed-effects binomial regression model with prompt type, prior rate, and noise level as independent variables and expression typed (zero anaphor or pronoun) as the dependent variable. The model further included an interaction term between prompt type and noise level, random intercepts for participants and random slopes for prompt type by participants, as justified by the model comparison.

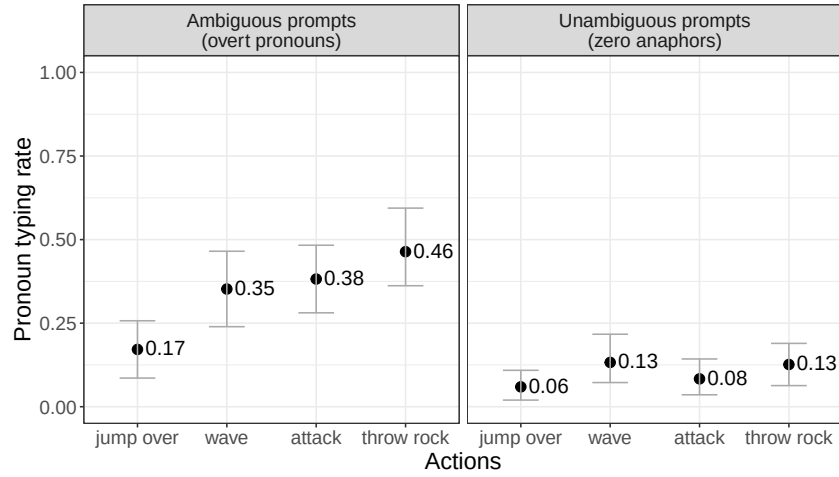


Figure 10: Experiment 2a (full dataset). Pronoun typing rate by action and prompt type.

	β	SE	z	p
(Intercept)	-14.30	2.15	-6.66	< 0.001
ambiguous prompt	10.63	1.97	5.39	< 0.001
high noise	2.74	1.26	2.18	0.030
action prior	3.88	1.05	3.71	< 0.001
ambiguous prompt : high noise	-3.18	1.33	-2.40	0.017

Table 4: Experiment 2a (full dataset). Expression typed (zero anaphor or pronoun). Model syntax: `glmer(expression ~ prompt type * noise level + action prior + (1 + prompt type | subject id))`. Unambiguous prompts and low noise serve as reference levels.

Recall that for each trial, participants chose the event outcome that they believed was described in the prompt. Unambiguous prompts were only compatible with the outcome where the agent of the first sub-event falls down. If participants selected a scene where the patient falls down, they might have misinterpreted the prompt. We therefore expected them to type a pronoun (that was originally not in the prompt) more often when they selected the patient outcomes. The link between outcome selection and pronoun reconstruction was strong: within trials containing unambiguous zero-anaphor prompts, participants typed a pronoun on 66% of trials in which they selected the patient-falling scene, compared with 9% of trials in which they selected the agent-falling scene.

Furthermore, if prior expectations over event outcomes affected prompt interpretation, we would expect insertions to occur more frequently for events with stronger patient bias. Indeed, participants in our experiment were more likely to report hearing a pronoun for actions with stronger patient bias, which surfaced in the analysis as a significant effect of action prior (Table 4, fourth row). The analysis also showed a significant effect of noise level for unambiguous prompts (Table 4, third row), with participants more likely to reconstruct a pronoun for high-noise prompts. Contrary to our predictions, the analysis showed no significant interaction between action priors and noise level (Table B.14, last row), however, it is possible that our dataset did not have sufficient sample size to detect such an interaction.

6.1.3. Discussion

The results of Experiment 2a show that expectations about event outcomes influenced listeners’ interpretation of anaphoric references in the prompt. The stronger the patient bias for an action, the more likely listeners were to interpret the anaphoric reference as referring to the patient. For unambiguous prompts (zero anaphors), the only grammatically valid interpretation was one in which the agent fell. However, listeners sometimes behaved as though the utterance contained a pronoun and interpreted it as referring to the patient. Consistent with the noisy-

channel model (Gibson et al., 2013; Levy, 2008), for unambiguous prompts, high noise increased both patient-outcome selection and pronoun typing, suggesting that listeners were more likely to favor a plausible alternative interpretation when the speech signal was less reliable. The typed responses provide evidence consistent with pronoun reconstruction, although they do not distinguish perceptual insertion from later memory- or response-level reconstruction.

We observed that in Experiment 2a, participants typed the pronoun on only approximately 35% of trials in which it was present in the noisy prompt. Such a low rate can be an artifact of a specific prompt, possibly reflecting the fact that in the sequence “and it”, both words end with alveolar plosive sounds, which, in the presence of noise, can be perceived as a single prolonged sound. Experiment 2b tests how accurate participants will be without the auditory interference and whether the effects observed in Experiment 2a will be reproduced in the clean audio condition.

6.2. Experiment 2b: Speech perception without added auditory noise

In Experiment 2a, we examined how participants interpret referring expressions in noisy conditions. To establish a baseline and test whether prior effects persist even when the speech signal is presented without added auditory noise, we conducted a follow-up experiment using clean audio.⁹ The analysis reported for this experiment is exploratory.

6.2.1. Methods

Experiment 2b followed the same design as Experiment 2a, with the primary difference that all audio files were presented without noise. Participants completed the same two tasks: after listening to the description, they selected the corresponding scene and typed the final part of the phrase they had heard. As in the main experiment, they evaluated four trials (one trial per action) in a random order, with prompt type and monster colors also randomized in each trial.

Hypotheses. Experiment 2b served as a baseline comparison: the noisy-channel model predicts that without noise, listeners would largely correctly identify what they heard. For unambiguous prompts, we expected participants to primarily select the scene where the agent falls, as the alternative scene would be incompatible with the audible prompt. In contrast, ambiguous prompts allow for both outcomes, so we predicted that listeners’ prior expectations would influence how often they select the scene where the patient falls.

⁹The experiment is available via the link:
<https://www.cognitive-modeling-experiments.uni-tuebingen.de/publix/6E0VsejEqTp>.

Participants. Because Experiment 2b included only one auditory condition rather than the two noisy conditions used in Experiment 2a, we recruited 100 participants to obtain approximately the same expected number of observations per action-by-prompt-type cell as in each noise condition of Experiment 2a. This target was pragmatic rather than based on a separate power analysis. After excluding 3 non-native speakers and trials where the typed response contained neither a pronoun nor a zero anaphor (which led to the exclusion of 5 more participants who had no usable trials), a total of 366 trials from 92 participants were included in the analysis. Out of 366 trials, 174 featured ambiguous prompts and 192 unambiguous ones.

6.2.2. Results

Overall, as we expected, participants were largely accurate in reporting what they heard: when the pronoun was present in the prompt, they typed it in their responses 83% of the time, and when the pronoun was absent, they only reported hearing it 4% of the time. This result provides a sanity check that participants pay attention to the linguistic prompt and perform the task as expected.

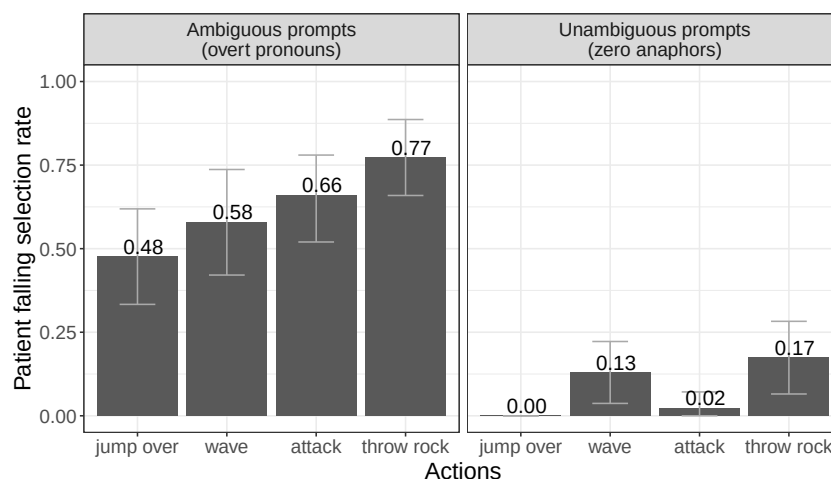


Figure 11: Experiment 2b (full dataset). Selection rate of the patient-falling scene by action and prompt type.

Analysis of the selected outcomes. Action priors influenced event selection across both prompt types (Figure 11). Although participants rarely chose the patient scene in the unambiguous condition without noise, they were more likely to select the patient-falling outcome for actions with stronger patient bias. Mixed-effects

binomial regression models with random intercepts for participants and items¹⁰ revealed significant main effects of prompt type and priors on event selection outcomes (Table 5) and no significant interaction between prompt type and priors (Table C.15).

	β	SE	z	p
(Intercept)	-5.81	0.90	-6.44	< 0.001
ambiguous prompt	3.84	0.48	7.96	< 0.001
action prior	3.93	0.99	3.98	< 0.001

Table 5: Experiment 2b (full dataset). Selected outcome (agent or patient falling). Model syntax: `glmer(selected outcome ~ prompt type + action prior + (1 | subject id) + (1 | item))`. Unambiguous prompts serve as the reference level.

The overall rate at which listeners selected the patient scene in the ambiguous case was higher than in Experiment 2a (compare Figure 11 with Figure 9, left panels), likely because participants often missed the pronouns when prompts were masked with noise, making the patient option unavailable, and consequently, it was selected less frequently.

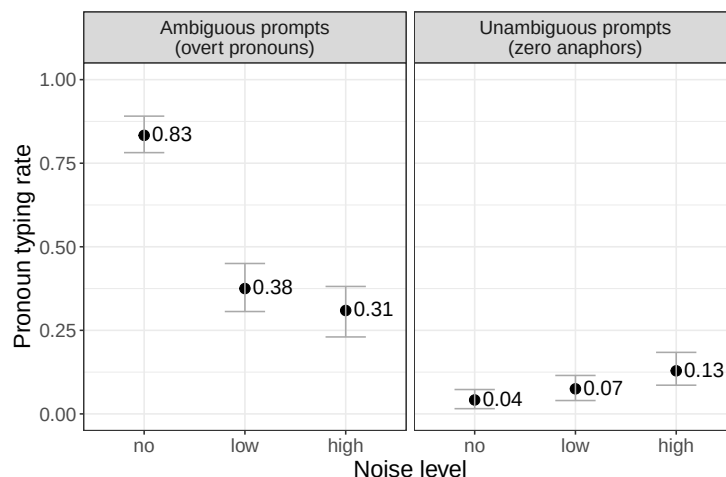


Figure 12: Experiments 2a and 2b combined. Pronoun typing rate by noise level and prompt type.

Analysis of the expressions typed. In Experiment 2b, participants correctly typed the pronoun when it was present 83% of the time and correctly omitted it when

¹⁰The random-effects structure here and in the next section mirrored that of the models of the main experiment with noise.

absent 96% of the time. In Experiment 2a these figures were noticeably lower: participants correctly typed a pronoun when it was present in the prompt 38% of the time in the low-noise condition and 31% in the high-noise condition and typed no pronoun when they heard unambiguous prompts 93% and 87% of the time respectively (Figure 12).

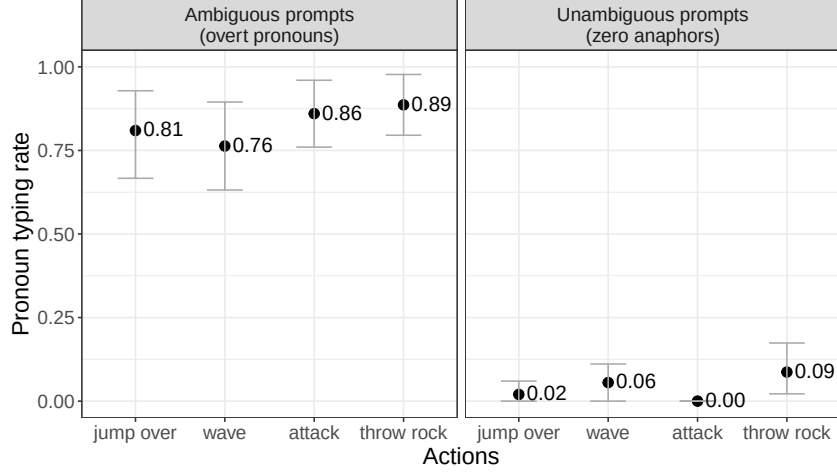


Figure 13: Experiment 2b (full dataset). Pronoun typing rate by action and prompt type.¹¹

	β	SE	z	p
(Intercept)	-71.02	15.15	-4.69	< 0.001
ambiguous prompt	76.68	15.37	4.99	< 0.001
action prior	61.24	15.23	4.02	< 0.001
ambiguous prompt : action prior	-53.68	15.92	-3.37	0.001

Table 6: Experiment 2b (full dataset). Expression typed (zero anaphor or pronoun). Model syntax: `glmer(expression ~ prompt type * action prior + (1 + prompt type | subject id))`. Unambiguous prompts serve as the reference level.

Notably, prior-related biases persisted in the unambiguous condition even without added auditory noise (Figure 13). A mixed-effects binomial regression model with random intercepts for participants and slopes for prompt type by participants revealed a significant interaction between prompt type and priors on typed

¹¹In the right panel (unambiguous case) of this figure and Figure 11, the *attack* action shows a lower value than *wave*, which appears to contradict the priors obtained in the norming study. However, the priors for these two actions are both mid-range and very close to each other, so some random variation in their relative values across experiments is to be expected. Such variance can be particularly high when the values are as low as in this case.

responses (Table 6, last row). It shows a significant effect of priors in the unambiguous case, the reference level in the model (Table 6, third row). However, in the ambiguous case, the effect of priors on typed expression was not significant (see Table C.16, where the reference level was changed to ambiguous prompts to test the simple effect of action priors in this case).

6.2.3. Discussion

Overall, when no auditory noise was added, participants were largely accurate in identifying the content of the prompt. Nevertheless, their patient outcome selection rates for both ambiguous and unambiguous prompts were still influenced by prior expectations. This effect is particularly notable in the unambiguous case with zero anaphors in the prompt, where the patient selection and pronoun insertion rates are very low, since such interpretations are, strictly speaking, errors. Yet both rates remain affected by action priors.

Our predictions were supported in the ambiguous case, where we see the effect of priors on the selected outcome, but not on the typed response. The presence of a pronoun is compatible with both outcomes, so there is no pressure to omit the pronoun in order to select the preferred interpretation. On the other hand, in the unambiguous case, there is such pressure. And indeed, participants sometimes selected and typed responses consistent with representing the utterance as though a pronoun had been present, making the event description more plausible, even when no auditory noise was added. These results suggest that even a very reliable linguistic signal does not entirely suppress prior-driven interpretation: event plausibility biases interpretation even when the auditory signal is comparatively clear.

7. General discussion

Information predictability and linguistic predictability often pattern together, making it difficult to establish a precise relation between the choices speakers make and their underlying causes. This difficulty might have partially contributed to mixed findings regarding the role of predictability reported in literature. In the current work, we investigated the contribution of information predictability, operationalized as event predictability, and linguistic constraints on language production and interpretation. We focused on the phenomenon of anaphoric reference, a domain where the two sources of knowledge are typically intertwined: the same structural, thematic, and discourse-level factors that shape which referent is linguistically salient also correlate with which event outcomes are world-knowledge-plausible.

We built on the artificial-world paradigm (Stegemann-Philipps et al., 2021), creating our own material. The material depicted imaginary monsters, who first

engaged in one of four interactions (*jump over*, *wave*, *attack*, or *throw rock*) followed by an outcome event in which either the agent or the patient of the initial interaction fell down. In this setup, speakers did not have to decide which referent to mention next but only how to refer to that referent. The artificial-world paradigm additionally reduced the influence of lexical co-occurrence and linguistic next-mention probability on the choice of referring expressions, because prior expectations were assessed and manipulated through animations without linguistic event descriptions.

This paradigm allowed us to estimate both the effect of natural action priors and priors over event outcomes induced in the course of the speech production experiment. While the effect of action priors was more permanent and present both in the production and comprehension experiments reported in this paper, in the production part, we experimentally intervened on event predictability, strengthening evidence for the role of predictability in communication. Furthermore, the manipulation of priors over event outcomes in the production study enabled us to disentangle the contributions of the antecedent position and event predictability.

The production results showed that event-level expectations influenced referring expression choice alongside established linguistic constraints. In Experiment 1a, speakers produced more overt referring expressions for surprising outcomes dependent on both action prior bias and trained outcome bias. Experiment 1b indicated that the larger first-to-second-block change in Experiment 1a was not fully explained by general progression through the task: changes also occurred without training, but the first-to-second-block contrast was significantly larger following training. The comparison of the first and third blocks of trials in Experiment 1a, featuring exactly the same events, showed that the effect of experiment-induced priors was transient. After participants' expectations over monster strength were disrupted by surprising events in the second block, they no longer relied on their trained priors. What remained was the effect of trial order, as shown by the analysis of Experiment 1b: the difference between the first and third blocks was similar in both experiments, with and without training.

Our data further revealed that linguistic accessibility also exerted a strong influence on production. Speakers used less overt forms more often when the antecedent had appeared as the subject and agent of the preceding event than when it had appeared as the object and patient. Participants overwhelmingly produced definite NPs to refer to patient antecedents, making the contrast between familiar and surprising events rather subtle in that case. Overall, the effect of antecedent position replicates the established preference to use reduced forms to refer to prominent referents (Arnold, 2001, 2010; Brennan, 1995) in a visually grounded free-production paradigm in which the next-mentioned referent was fixed. Because subjecthood, thematic role, order of mention, and likely topicality were aligned in

our materials, the experiment cannot determine which of these properties was responsible for the contrast. It nevertheless shows that event predictability affected referential choice in the presence of a substantial structural effect.

The utterance production patterns also indicated that speakers not only preferred definite NPs over pronouns when referring to surprising referents but also produced less zero anaphors than for familiar referents. One possibility is that speakers anticipated listeners reinterpreting a zero anaphor as though a pronoun had been present, thereby making a more plausible patient interpretation available. Our production experiments did not directly demonstrate that speakers explicitly anticipated this particular reconstruction process, but the comprehension experiments showed that listeners sometimes produced responses consistent with representing a zero-anaphor utterance as though it contained a pronoun if that misinterpretation made the event description more plausible. In line with the noisy-channel model of speech comprehension (Levy, 2008), under uncertainty listeners may have inferred that a short pronoun such as *it* was present in the intended utterance, even though it was absent from the presented signal. Such an insertion carries a relatively low edit cost compared with replacing a lexically specific definite NP (Gibson et al., 2013; Levy, 2008). By contrast, reconstructing an utterance containing a definite NP as referring to the other character would require replacing a longer, lexically specific expression and should incur a greater edit cost.

Experiment 2b helped us isolate the role of auditory noise itself. Descriptively, participants identified the referring expression more accurately without added auditory noise, and pronoun omissions were less frequent than in Experiment 2a. Action-prior effects, however, did not go away: outcome plausibility still shaped which event outcome participants chose, even for unambiguous prompts. Within Experiment 2a, higher noise increased responses that conflicted with unambiguous prompts. What remains unclear is how much of this residual effect reflects other sources of noise in the system, such as general perceptual imprecision, memory, attention, or decision-level factors, rather than the auditory channel specifically.

It further remains to be assessed how the results that we obtained by evaluating only four types of interactions generalize to other types of actions. Additionally, the experiments reported in this work were underpowered relative to the preregistered sample-size targets, limiting our ability to assess some of the interaction effects. Participant and trial exclusions resulted in smaller final samples than anticipated by the power analyses, and these reductions were not accounted for. This limitation does not undermine the effects we reported, but it means that null results, most notably the absence of an interaction between action priors and noise, should be interpreted cautiously, as they may reflect insufficient statistical power.

Taken together, the production and comprehension experiments provide con-

verging evidence that expectations over event outcomes influence both sides of referential communication. In production, surprising events were associated with greater referential overttness. In comprehension, event plausibility guided the interpretation of ambiguous pronouns and sometimes biased listeners’ responses toward a more plausible event interpretation that was incompatible with the presented zero-anaphor form. These results are consistent with an audience-oriented, RSA-style account in which speakers’ choices are sensitive to listeners’ interpretations. Experiments 2a and 2b make an audience-oriented explanation more plausible by showing that listeners do sometimes favor more plausible alternative interpretations of reduced utterances. Whether participants execute RSA-like recursions online or have adapted a more general strategy to increase the likelihood of choosing more specific utterance when expecting higher uncertainty in the listener remains to be examined further.

In sum, this paper contributes to the understanding of referring expression choice as sensitive to event predictability. Event-level priors showed a quantifiable association with both the production and interpretation of anaphoric references, over and above the linguistic predictors included in the models. Crucially, parallel effects of action-level expectations were observed in the two domains: speakers chose more explicit forms when describing less likely events; whereas listeners were more likely to (mis)interpret utterances in favor of more plausible event descriptions—even when the linguistic signal was unambiguous. These findings extend noisy-channel reasoning to anaphoric interpretation and show that referential communication depends not only on the linguistic accessibility of discourse referents, but also on expectations about how events are likely to unfold.

8. Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project number 519109539. Martin Butz acknowledges funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—381713393, 467045002, 564829065. He is a member of the Machine Learning Cluster of Excellence—90727645. We would like to thank Christian Stegemann-Philipps for initiating this project and for many fruitful discussions. Lena Holzwarth created the original artwork and prototyped the animations for stimuli creation. We would further like to thank Marjolein von Os for sharing her expertise on speech perception under noise. The graphical abstract for this paper was co-designed with the artist Tanja Russita. Tino Sering and Robin Hörnig shared helpful suggestions on the statistical analysis. We are grateful to the audiences of XPrag, Sinn und Bedeutung, and CogSci for their helpful feedback. Finally, we truly appreciate the thoughtful and constructive feedback that

anonymous reviewers and editors generously shared with us. All errors remain our own.

Appendix A. Experiment 1a: Speech production with training

Appendix A.1. Deviations from preregistration

In our preregistered analysis, we planned to evaluate predictability separately for subject (agent) and non-subject (patient) antecedents, focusing on the contrast between familiar and surprising trials. However, this type of analysis does not allow us to estimate the effect of antecedent position on the production of referring expressions, as well as the potential interaction between predictability and antecedent position.¹² Furthermore, we planned to treat the type of referring expression as a binomial dependent variable grouping together two reduced forms of reference: zero anaphors and pronouns. These forms, however, carry different properties, which could potentially affect speakers’ choices. Pronouns are at least in theory ambiguous, even if pronouns show a subject bias. Zero anaphors are unambiguous, but due to inattention or noise, listeners may misperceive the utterance as containing a pronoun.

We made the following changes to the analysis plan to minimize data post-processing and make the analysis more conservative. We also added exploratory analyses to the original preregistration. For transparency, we report the originally planned analyses in Appendix A.2.

1. Changes to the preregistration: effect of predictability
 - (a) The referring expression produced is treated not as a binary variable (noun phrase or reduced form) but as a three-way ordinal dependent variable with the levels “zero anaphor”, “pronoun”, and “definite NP”. The levels are ordered by their phonetic overttness. The most overt form (a definite NP) is less phonetically and referentially ambiguous.
 - (b) We perform the analysis on the full dataset including both agent- and patient-antecedent cases.
2. Exploratory analysis
 - (a) In addition to the effect of predictability, we estimate the effect of antecedent position.
 - (b) In addition to the difference between familiar and surprising events, we estimate the difference between familiar and post-surprising events.
 - (c) We estimate the effect of action priors, as well as the interaction between action priors, event predictability, and antecedent position.

¹²We would like to thank an anonymous reviewer for highlighting this limitation.

Appendix A.2. Preregistered data analysis

In this analysis, we assessed whether event predictability affects the type of anaphoric references produced. More specifically, we asked whether participants were more likely to produce a definite NP, compared to a reduced form of reference (a pronoun or a zero anaphor). We grouped zero anaphors and pronouns together since they both are associated with interpretation uncertainty: pronouns feature referential ambiguity, while zero anaphors can potentially be misinterpreted. In contrast, definite NPs feature unambiguous descriptions and are more phonologically overt, making misinterpretation less likely.

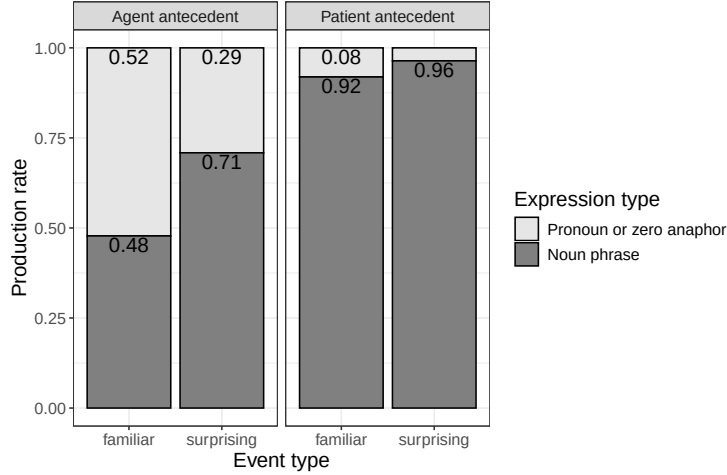


Figure A.14: Experiment 1a (the first four target trials). Produced referring expressions (noun phrase or reduced form) by event type (familiar vs. surprising) and antecedent position (agent vs. patient).

We analyzed the data using a mixed-effects binomial regression model. Event type (familiar vs. surprising) served as a treatment-coded independent variable with familiar events serving as the reference level and produced expression type (NP vs. reduced) as the binary outcome variable. We included only the first four target trials in the analysis, ignoring post-surprising events, as specified in the preregistered plan. The model included random intercepts for participants.¹³

We performed separate analyses for agent and patient antecedents. For the agent-antecedent case, speakers produced more noun phrases (71%) for surprising events compared to familiar events (48%) (Figure A.14, left panel). This difference

¹³Model syntax: `glmer(bin expression ~ event type + (1 | subject id))`. Random slopes for event type by participant were not included because the number of random effects would have exceeded the number of observations. Random intercepts for items were excluded due to zero variance, which resulted in a singular fit.

was statistically significant ($\beta = 1.47, SE = 0.40, z = 3.69, p < 0.001$), supporting our hypothesis that event predictability affects the choice of referring expressions. In contrast, in the patient-antecedent condition (Figure A.14, right panel) the effect of predictability was not significant ($\beta = 0.86, SE = 0.63, z = 1.36, p = 0.174$). This null result may be attributed to limited statistical power, as the exclusion of a substantial number of participants reduced the sample to 139, corresponding to approximately 60% power to detect the effect.

Appendix A.3. Exploratory data analysis

	β	SE	z	p
event (familiar)	-1.48	0.31	-4.72	< 0.001
event (post-surprising)	-0.25	0.30	-0.83	0.407
action priors	5.00	1.19	4.19	< 0.001
patient	11.13	2.73	4.08	< 0.001
action priors : patient	-6.75	3.36	-2.01	0.045

Table A.7: Experiment 1a (full dataset). Produced referring expression (zero anaphor, pronoun, or NP). Model syntax: `clmm(expression ~ relevel(event type, ref = "surprising") + action prior * antecedent + (1 + antecedent | subject id) + (1 | item))`. Surprising events and agent antecedents serve as reference levels.

	β	SE	z	p
event (surprising)	1.48	0.31	4.72	< 0.001
event (post-surprising)	1.23	0.30	4.04	< 0.001
action prior	-1.75	3.21	-0.55	0.585
agent	-11.13	2.73	-4.08	< 0.001
action priors : agent	6.75	3.36	2.01	0.045

Table A.8: Experiment 1a (full dataset): produced referring expression (zero anaphor, pronoun, or NP). Model syntax: `clmm(expression ~ event type + action prior * relevel(antecedent, ref = "patient") + (1 + antecedent | subject id) + (1 | item))`. Familiar events and patient antecedents serve as reference levels.

		<i>diff</i>	<i>SE</i>	<i>z</i>	<i>p</i>
zero anaphor	jump over - wave	0.183	0.051	3.61	0.002
	jump over - attack	0.208	0.057	3.68	0.001
	jump over - throw rock	0.319	0.073	4.38	< 0.001
	wave - attack	0.026	0.006	4.13	< 0.001
	wave - throw rock	0.136	0.026	5.24	< 0.001
	attack - throw rock	0.111	0.021	5.40	< 0.001
pronoun	jump over - wave	0.043	0.014	2.98	0.015
	jump over - attack	0.055	0.017	3.21	0.007
	jump over - throw rock	0.141	0.034	4.21	< 0.001
	wave - attack	0.013	0.004	3.20	0.008
	wave - throw rock	0.099	0.024	4.04	< 0.001
	attack - throw rock	0.086	0.021	4.16	< 0.001
noun phrase	jump over - wave	-0.225	0.053	-4.27	< 0.001
	jump over - attack	-0.264	0.062	-4.28	< 0.001
	jump over - throw rock	-0.460	0.092	-5.03	< 0.001
	wave - attack	-0.038	0.009	-4.29	< 0.001
	wave - throw rock	-0.235	0.042	-5.66	< 0.001
	attack - throw rock	-0.197	0.033	-5.95	< 0.001

Table A.9: Experiment 1a. Tukey-adjusted pairwise comparisons of model-predicted marginal probabilities by action within each expression type. Contrasts were obtained from the cumulative link mixed model (Table 1), evaluated for agent antecedents averaged across event types.

Appendix B. Experiment 2a: Speech perception under noise

Appendix B.1. Preregistered power analysis

To estimate the required number of participants, we performed a simulation-based power analysis using the *simr* package in R and the data from the pilot study (Achimova et al., 2024). For the simulation, we adjusted the design matrix by adding one further type of action, included action priors from the norming study as a continuous predictor, and adjusted the results to account for the reduced number of trials. Following the preregistration, for ambiguous prompts, we estimated the power needed to detect the effect of action priors on the selected outcome,¹⁴ while for unambiguous prompts, we were concerned with the effect of action priors, as well as the interaction of action priors with noise level, on both dependent variables: selected outcome and typed expression.¹⁵

¹⁴Model syntax: `glmer(selected outcome ~ action prior + (1 + action prior | subject id))`.

¹⁵Model syntax: `glmer(selected outcome ~ action prior + (1 + action prior | subject id))`, `glmer(expression ~ action prior + (1 + action prior | subject id))`, `glmer(selected outcome ~`

The analysis revealed that a sample size of 200 participants would be sufficient to detect all the considered effects, but this estimate contained an error,¹⁶ that led to an underestimation of the required number of participants. However, since we switched to the analysis of the whole dataset (ambiguous and unambiguous prompts together), and due to a larger effect size than expected, we were able to detect the effect of action priors but not the interaction with noise level. The full power estimation protocol is available at https://osf.io/nr9py/files?view_only=0db29dd5eeeb4f6f943ae22292418e4f

Appendix B.2. Preregistered data analysis

In line with the preregistration, we conducted separate analyses for ambiguous and unambiguous prompts, as our hypotheses differed across these two subsets. For ambiguous prompts, we examined whether action priors influenced the selected outcome (i.e., whether the agent or the patient was judged to fall). We analyzed the data using a mixed-effects binomial regression model, with action prior as the independent variable and the selected outcome as the binomial dependent variable. The model included random intercepts for both participants and items.¹⁷ The analysis revealed a significant effect of action priors on outcome selection ($\beta = 4.67, SE = 1.11, z = 4.22, p < 0.001$).

For the unambiguous prompts, we additionally examined insertion rates, i.e., the typed expression dependent variable. We expected both dependent variables to be influenced not only by action priors but also by noise level. Crucially, we were interested in the interaction between action priors and noise level, predicting that the effect of action priors would be stronger in the high-noise condition.

For each dependent variable, we fitted a mixed-effects binomial regression model including action prior and noise level as independent variables, along with their interaction. The models included random intercepts for participants and items, as well as random slopes for noise level by participants, as justified by the model comparison (Tables B.10 and B.11). As the interaction was not significant for either the selected outcome or the typed expression and did not improve model fit according to the model comparison, we refitted the models without the interaction terms to assess the main effects of action priors and noise level (Tables B.12

action prior * noise level + (1 + action prior * noise level | subject id)), and `glmer(expression ~ action prior * noise level + (1 + action prior * noise level | subject id))`.

¹⁶Adding an additional value of a continuous variable (action priors) worked incorrectly in the *simr* package. Also, the change in the number of trials per participant was not correctly accounted for.

¹⁷Model syntax: `glmer(selected outcome ~ action prior + (1 | subject id) + (1 | item))`. Random slopes for action prior by participants were not included because doing so would have resulted in more random-effects parameters than observations.

and B.13). The results showed that both factors significantly affected the selected outcome, but neither had a significant effect on the typed expression.

	β	SE	z	p
(Intercept)	-14.38	3.94	-3.65	< 0.001
action prior	6.64	3.95	1.68	0.093
high noise	10.52	4.05	2.60	0.009
action prior : high noise	-3.73	4.15	-0.90	0.368

Table B.10: Experiment 2a (unambiguous prompts). Selected outcome (agent or patient falling). Interaction model. Model syntax: `glmer(selected outcome ~ action prior * noise level + (1 + noise level | subject id))`. Low noise as the reference level.

	β	SE	z	p
(Intercept)	-12.22	4.97	-2.46	0.014
action prior	0.98	6.09	0.16	0.873
high noise	0.04	5.73	0.01	0.994
action prior : high noise	1.41	7.19	0.20	0.844

Table B.11: Experiment 2a (unambiguous prompts). Expression typed (zero anaphor or pronoun). Interaction model. Model syntax: `glmer(expression ~ action prior * noise level + (1 + noise level | subject id) + (1 | item))`. Low noise as the reference level.

	β	SE	z	p
(Intercept)	-11.54	1.77	-6.50	< 0.001
action prior	3.33	1.17	2.84	0.005
high noise	7.37	1.53	4.81	< 0.001

Table B.12: Experiment 2a (unambiguous prompts). Selected outcome (agent or patient falling). Main effects. Model syntax: `glmer(selected outcome ~ action prior + noise level + (1 + noise level | subject id))`. Low noise as the reference level.

	β	SE	z	p
(Intercept)	-12.94	3.65	-3.55	< 0.001
action prior	1.96	3.50	0.56	0.576
high noise	1.06	2.62	0.41	0.685

Table B.13: Experiment 2a (unambiguous prompts). Expression typed (zero anaphor or pronoun). Main effects. Model syntax: `glmer(expression ~ action prior + noise level + (1 + noise level | subject id))`. Low noise as the reference level.

Appendix B.3. Exploratory data analysis

	β	SE	z	p
(Intercept)	-16.09	3.53	-4.56	< 0.001
ambiguous prompt	12.48	3.46	3.61	< 0.001
high noise	3.20	2.07	1.55	0.122
action prior	5.94	3.26	1.82	0.068
ambiguous prompt : high noise	-3.26	1.35	-2.42	0.016
ambiguous prompt : action prior	-2.11	3.12	-0.68	0.499
action prior : high noise	-0.54	2.17	-0.25	0.803

Table B.14: Experiment 2a (full dataset). Expression typed (zero anaphor or pronoun). Pairwise interaction model. Model syntax: `glmer(expression ~ prompt type * (noise level + action prior) + noise level * action prior + (1 + prompt type | subject id))`. Unambiguous prompts and low noise serve as reference levels.

Appendix C. Experiment 2b: Speech perception without noise

	β	SE	z	p
(Intercept)	-6.67	1.42	-4.69	< 0.001
ambiguous prompt	5.11	1.62	3.15	0.002
action prior	5.09	1.74	2.93	0.003
ambiguous prompt : action prior	-1.76	2.12	-0.83	0.406

Table C.15: Experiment 2b (full dataset). Selected outcome (agent or patient falling). Interaction model. Model syntax: `glmer(selected outcome ~ prompt type * action prior + (1 | subject id) + (1 | item))`. Unambiguous prompts serve as the reference level.

	β	SE	z	p
(Intercept)	5.66	2.70	2.10	0.036
unambiguous prompt	-76.54	15.69	-4.88	< 0.001
action priors	7.56	4.97	1.52	0.128
unambiguous prompt : action prior	53.54	16.19	3.31	0.001

Table C.16: Experiment 2b (full dataset). Expression typed (zero anaphor or pronoun). Model syntax: `glmer(expression ~ relevel(prompt type, ref = "ambiguous") * action prior + (1 + prompt type | subject id))`. Ambiguous prompts serve as the reference level.

References

- Achimova, A., van Os, M., Demberg, V., Butz, M.V., 2024. Interpreting implausible event descriptions under noise, in: Proceedings of the Annual Meeting of the Cognitive Science Society, pp. 3399–3406.
- Achimova, A., Stegemann-Philipps, C., Winkler, S., Butz, M.V., 2022. Anaphoric reference in descriptions of surprising events, in: Proceedings of Sinn und Bedeutung, pp. 20–34.
- Arnold, J.E., 1998. Reference form and discourse patterns. Stanford University.
- Arnold, J.E., 2001. The effect of thematic roles 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, 495–527.
- Arnold, J.E., 2010. How speakers refer: The role of accessibility. *Language and Linguistics Compass* 4, 187–203.
- Arnold, J.E., Griffin, Z.M., 2007. The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language* 56, 521–536.
- 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, 31–56.
- Baldwin, D.A., Kosie, J.E., 2021. How does the mind render streaming experience as events? *Topics in Cognitive Science* 13, 79–105.
- Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67, 1–48.
- Benz, A., 2012. Errors in pragmatics. *Journal of Logic, Language and Information* 21, 97–116.
- Bergen, L., Goodman, N.D., 2015. The strategic use of noise in pragmatic reasoning. *Topics in Cognitive Science* 7, 336–350.
- Boothroyd, A., Nitttrouer, S., 1988. Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America* 84, 101–114.

- Bott, O., Solstad, T., 2023. The production of referring expressions is influenced by the likelihood of next mention. *Quarterly Journal of Experimental Psychology* 76, 2256–2284.
- Brennan, S.E., 1995. Centering attention in discourse. *Language and Cognitive Processes* 10, 137–167.
- Brochhagen, T., 2020. Signalling under uncertainty: Interpretative alignment without a common prior. *British Journal for the Philosophy of Science* 71, 471–496.
- Butz, M.V., 2017. Which Structures are Out There?: Learning Predictive Compositional Concepts Based on Social Sensorimotor Explorations. Johannes Gutenberg-Universität Mainz.
- Butz, M.V., Achimova, A., Bilkey, D., Knott, A., 2021. Event-predictive cognition: A root for conceptual human thought. *Topics in Cognitive Science* 13, 10–24.
- Caramazza, A., Grober, E., Garvey, C., Yates, J., 1977. Comprehension of anaphoric pronouns. *Journal of Verbal Learning and Verbal Behavior* 16, 601–609.
- Christensen, R.H.B., 2026. ordinal—Regression Models for Ordinal Data. URL: <https://CRAN.R-project.org/package=ordinal>. R package version 2026.7-26.
- Davies, C., Arnold, J.E., 2019. Reference and informativeness, in: *The Oxford Handbook of Experimental Semantics and Pragmatics*. Oxford University Press, pp. 474 – 493. doi:10.1093/oxfordhb/9780198791768.013.6.
- Demberg, V., Kravtchenko, E., Loy, J.E., 2023. A systematic evaluation of factors affecting referring expression choice in passage completion tasks. *Journal of Memory and Language* 130, 104413.
- Elman, J.L., Kehler, A., Rohde, H., 2006. Event structure and discourse coherence biases in pronoun interpretation, in: *Proceedings of the Annual Meeting of the Cognitive Science Society*, pp. 697–702.
- Frank, M.C., Goodman, N.D., 2012. Predicting pragmatic reasoning in language games. *Science* 336, 998–998.
- Franke, M., Jäger, G., 2016. Probabilistic pragmatics, or why Bayes’ rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft* 35, 3–44.

- Fukumura, K., Van Gompel, R.P., 2010. Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language* 62, 52–66.
- Garvey, C., Caramazza, A., 1974. Implicit causality in verbs. *Linguistic Inquiry* 5, 459–464.
- Garvey, C., Caramazza, A., Yates, J., 1974. Factors influencing assignment of pronoun antecedents. *Cognition* 3, 227–243.
- Gibson, E., Bergen, L., Piantadosi, S.T., 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences* 110, 8051–8056.
- Goodman, N.D., Frank, M.C., 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences* 20, 818–829.
- Green, P., MacLeod, C.J., 2016. simr: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution* 7, 493–498. URL: <https://CRAN.R-project.org/package=simr>, doi:10.1111/2041-210X.12504.
- Grice, H.P., 1975. Logic and conversation, in: *Speech Acts*. Brill, pp. 41–58.
- Grosz, B.J., Joshi, A., Weinstein, S., 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21, 203–225.
- Guan, S., Arnold, J.E., 2021. The predictability of implicit causes: Testing frequency and topicality explanations. *Discourse Processes* 58, 943–969.
- Hannemann, R., Obleser, J., Eulitz, C., 2007. Top-down knowledge supports the retrieval of lexical information from degraded speech. *Brain Research* 1153, 134–143.
- Hobbs, J.R., Stickel, M.E., Appelt, D.E., Martin, P., 1993. Interpretation as abduction. *Artificial Intelligence* 63, 69–142.
- Holler, A., Suckow, K., 2016. How clausal linking affects noun phrase salience in pronoun resolution. *Empirical Perspectives on Anaphora Resolution* , 61–85.
- Hwang, H., Lam, S.Y., Ni, W., Ren, H., 2022. The role of grammatical role and thematic role predictability in reference form production in Mandarin Chinese. *Frontiers in Psychology* 13, 930572.

- Jäger, T.F., 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61, 23–62.
- Kaiser, E., Li, D.C.H., Holsinger, E., 2011. Exploring the lexical and acoustic consequences of referential predictability, in: *Discourse Anaphora and Anaphor Resolution Colloquium*, Springer. pp. 171–183.
- Kehler, A., 2002. Coherence, reference, and the theory of grammar. volume 380. CSLI publications Stanford, CA.
- 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.
- Kuperberg, G.R., 2021. Tea with milk? A hierarchical generative framework of sequential event comprehension. *Topics in Cognitive Science* 13, 256–298.
- Kurumada, C., Jäger, T.F., 2015. Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language* 83, 152–178.
- Lam, T.Q., Watson, D.G., 2010. Repetition is easy: Why repeated referents have reduced prominence. *Memory & Cognition* 38, 1137–1146.
- Lange, K., Kühn, S., Filevich, E., 2015. "Just another tool for online studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PloS One* 10, e0130834.
- Lee, S.H.y., Papafragou, A., 2024. Event-general conceptual categories organize verb semantics and acquisition cross-linguistically, in: *Proceedings of the Annual Meeting of the Cognitive Science Society*, pp. 4504–4511.
- Lenth, R.V., Piaskowski, J., 2026. emmeans: Estimated Marginal Means, aka Least-Squares Means. URL: <https://CRAN.R-project.org/package=emmeans>. R package version 2.0.4.
- Levesque, Hector, J., Davis, E., Morgenstern, L., 2012. The Winograd schema challenge, in: *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, pp. 552–561.

- Levy, R., 2008. A noisy-channel model of human sentence comprehension under uncertain input, in: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 234–243.
- Levy, R., Jäger, T.F., 2007. Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems (NIPS)* 19, 849–856.
- Liao, X., Brochhagen, T., Boleda, G., Mayol, L., 2025. The role of referent predictability in pronoun production: Insights from a Bayesian meta-analysis. *Language* 101, 445–499.
- Medina Fetterman, A.M., Vazquez, N.N., Arnold, J.E., 2022. The effects of semantic role predictability on the production of overt pronouns in Spanish. *Journal of Psycholinguistic Research* 51, 169–194.
- Miller, G.A., Heise, G.A., Lichten, W., 1951. The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology* 41, 329.
- Modi, A., Titov, I., Demberg, V., Sayeed, A., Pinkal, M., 2017. Modeling semantic expectation: Using script knowledge for referent prediction. *Transactions of the Association for Computational Linguistics* 5, 31–44.
- Obleser, J., Kotz, S.A., 2010. Expectancy constraints in degraded speech modulate the language comprehension network. *Cerebral Cortex* 20, 633–640.
- Papafragou, A., Hulbert, J., Trueswell, J., 2008. Does language guide event perception? Evidence from eye movements. *Cognition* 108, 155–184.
- Patterson, C., Schumacher, P.B., Nicenboim, B., Hagen, J., Kehler, A., 2022. A Bayesian approach to German personal and demonstrative pronouns. *Frontiers in Psychology* 12, 672927.
- Pickering, M.J., Majid, A., 2007. What are implicit causality and consequentiality? *Language and Cognitive Processes* 22, 780–788.
- R Core Team, 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: <https://www.R-project.org/>.
- Rohde, H., Kehler, A., 2014. Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience* 29, 912–927.

- Rosa, E.C., Arnold, J.E., 2017. Predictability affects production: Thematic roles can affect reference form selection. *Journal of Memory and Language* 94, 43–60.
- Sohoglu, E., Peelle, J.E., Carlyon, R.P., Davis, M.H., 2012. Predictive top-down integration of prior knowledge during speech perception. *Journal of Neuroscience* 32, 8443–8453.
- Stegemann-Philipps, C., Butz, M.V., Winkler, S., Achimova, A., 2021. Speakers use more informative referring expressions to describe surprising events, in: *Proceedings of the Annual Meeting of the Cognitive Science Society*, pp. 2897–2903.
- Stevenson, R.J., Crawley, R.A., Kleinman, D., 1994. Thematic roles, focus and the representation of events. *Language and Cognitive Processes* 9, 519–548.
- Ünal, E., Ji, Y., Papafragou, A., 2021. From event representation to linguistic meaning. *Topics in Cognitive Science* 13, 224–242.
- Van Os, M., Kray, J., Demberg, V., 2022. Rational speech comprehension: Interaction between predictability, acoustic signal, and noise. *Frontiers in Psychology* 13, 914239.
- Vogels, J., 2019. Both thematic role and next-mention biases affect pronoun use in Dutch, in: *Proceedings of the Annual Meeting of the Cognitive Science Society*, pp. 3029–3035.
- Weatherford, K.C., Arnold, J.E., 2021. Semantic predictability of implicit causality can affect referential form choice. *Cognition* 214, 104759.
- Winograd, T., 1972. Understanding natural language. *Cognitive Psychology* 3, 1–191.
- Ye, Y., Arnold, J.E., 2025. Implicit causality can affect pronoun use in fragment completion tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition* .
- Zacks, J.M., Speer, N.K., Swallow, K.M., Braver, T.S., Reynolds, J.R., 2007. Event perception: A mind-brain perspective. *Psychological Bulletin* 133, 273.
- Zerkle, S.A., Arnold, J.E., 2019. Does pre-planning explain why predictability affects reference production? *Dialogue & Discourse* 10, 34–55.
- Zipf, G.K., 1949. Human behavior and the principle of least effort: An introduction to human ecology. Addison-Wesley Press.