Comparing models of pronoun production and interpretation via observational and experimental evidence

Xixian Liao
Universitat Pompeu Fabra
0000-0002-8214-3659

Gemma Boleda
Universitat Pompeu Fabra
0000-0001-6140-7080

Hannah Rohde
University of Edinburgh
0000-0002-9356-3229

Laia Mayol
Universitat Pompeu Fabra
0000-0001-5386-816X

Abstract Pronouns like *she* are frequently produced by speakers to refer to entities in discourse. For communication to be successful, comprehenders must be able to interpret these pronouns by identifying the appropriate referent. In the existing literature, three main models of pronoun production and interpretation have been proposed. These models have traditionally been tested through story continuation tasks, using carefully designed stimuli. In our study, we take a different approach by utilizing naturalistic passages from corpora, in two analyses, one observational and one experimental. Our analyses support the Bayesian model. In this model and in our experimental data, the relationship between pronoun production and interpretation can be captured using Bayes' rule. Specifically, pronoun interpretation is affected both by the probability that the referent will be mentioned next and by the probability that a pronoun will be used to refer to that referent. Moreover, both observational and experimental data provide evidence that pronoun production biases are insensitive to a set of meaning-driven factors —here, discourse relations —which do affect pronoun interpretation, in line with the prediction of the so-called strong form of the Bayesian Model.

Keywords: pronoun production; pronoun interpretation; Bayes; corpus

1 Introduction

Pronouns, ubiquitous in discourse, are vital linguistic tools for referencing entities. They directly influence communicative success between speaker and listener and are crucial for efficient communication. Traditional approaches to pronouns have long held the assumption that there exists a unified concept of salience that governs both pronoun production (i.e., speakers' choice to use a pronoun *she* instead of a full noun phrase like *Marie Curie*) and pronoun interpretation (i.e., addressees' resolution of the pronoun's reference) and that pronoun production and interpretation are guided by the same set of contextual factors (e.g., Givón 1983; Ariel 1990; Gundel et al. 1993). According to these theories, speakers select pronouns to refer to entities they believe are highly activated in the addressees' cognitive state, and addressees, in turn, interpret these pronouns as referring to the most contextually salient entities. For instance, one linguistic feature that signals salience is grammatical position, where entities mentioned in the subject position are typically considered as more salient and topical (e.g., Crawley et al. 1990; Brennan 1995). Therefore, addressees tend to favor subjects as the likely referents of pronouns, and speakers tend to use pronouns for entities in the subject position rather than those in more oblique grammatical roles.

However, more recent investigations provide evidence for certain asymmetries between pronoun production and interpretation (e.g., Stevenson et al. 1994; Arnold 2001; Kehler et al. 2008; Mayol 2018), and some have proposed alternative models that relate pronoun production and interpretation in ways that go beyond mere mirroring of one another. One such model is the Bayesian Model, which posits that the relationship between pronoun interpretation and pronoun production can be captured using Bayesian principles (Kehler et al. 2008; Kehler & Rohde 2013). Specifically, pronoun interpretation is characterized as a combination of next-mention bias, which reflects the listener's estimate that a referent will get mentioned next, and pronoun production bias, which pertains to the listener's expectation that the speaker will use a pronoun to refer to that referent. This claim—that interpretation and production biases can be modelled through Bayes theorem—is known as the *weak* form of the Bayesian Model (e.g., Kehler & Rohde 2013; Rohde 2019; Zhan et al. 2020; Hoek et al. 2021; Patterson et al. 2022; see Section 2.2 for further details).

The *strong* form of this Bayesian model further posits that next-mention bias and pronoun production bias are influenced by distinct types of contextual factors. On the one hand, the factors conditioning the next-mention bias primarily stem from meaning-driven factors (e.g., verb type and discourse relations). On the other hand, the production bias is primarily influenced by factors that are grammatical (e.g., subjecthood) or related to information structure (specifically, topichood, which is inherently pragmatic in nature) while insensitive to the meaning-driven factors that are known to affect the next-mention bias.

Nevertheless, the existing literature presents conflicting evidence on this matter, as some studies suggest that both next-mention bias and pronoun production bias are influenced by meaning-driven factors like verb type and discourse relations (e.g. Arnold 2001; Rosa & Arnold 2017; Zerkle & Arnold 2019; Lindemann et al. 2020; Konuk & von Heusinger 2021; Weatherford & Arnold 2021; Medina Fetterman et al. 2022), while others find evidence to the contrary (e.g. Ferretti et al. 2009; Fukumura & Van Gompel 2010; Rohde & Kehler 2014; Rosa 2015; Holler & Suckow 2016; Mayol 2018; Kehler & Rohde 2019; Zhan et al. 2020; Frederiksen & Mayberry 2022; Hwang et al. 2022; Kravtchenko 2022; Lam & Hwang 2022; Patterson et al. 2022; Hwang 2023a).

To date, the Bayesian model, in both its weak and strong forms, has been primarily evaluated using experimental psycholinguistic methodologies (Rohde & Kehler 2014; Mayol 2018; Bader & Portele 2019; Kehler & Rohde 2019; Zhan et al. 2020; Patterson et al. 2022). However, there is a notable lack of naturally occurring language representation; the language production in prior experiments has predominantly been elicited using deliberately crafted contexts. Moreover, there's a limited empirical breadth since the majority of evidence we currently have is derived from studies centered on next-mention biases elicited by two specific verb types: transfer-of-possession verbs, such as *give* and *receive*, and implicit causality verbs, such as *surprise* and *admire*. Other domains, regrettably, have remained un(der)explored. These previously studied contexts have largely been confined to a simplistic world that encompasses a single event and two animate entities, as in *Mary received a book from Anna* or *John surprised Bill.* We have limited evidence to determine if the same biases can be elicited in passages describing more authentic, natural scenarios.

These limitations motivate our choice to utilize corpus data. We extract passages from coreference-annotated corpora that have been developed within the field of computational linguistics (Carlson et al. 2002; Weischedel et al. 2013). Using more naturalistic contexts, we contribute new empirical evidence to the generality of the Bayesian pronoun approach and to the ongoing debate on the sensitivity of pronoun production to meaning-driven factors. We also go beyond prior work that has focused on verb-driven effects in testing the role of meaning-driving factors by instead examining contexts that vary in the discourse relation that holds between sentences. It turned out that the constructions closely resembling the experimental stimuli with verb types (e.g., *John amazed Bob because*), as used in previous psycholinguistic experiments, were infrequent occurrences in our corpus texts. Instead, we examined discourse relations between clauses, specifically, Narration, Contrast, and Result, which represent more general semantic and pragmatic patterns in discourse. Through our analysis, we show that these relations, both when explicitly signaled by connectives and when manually identified by human annotators, exhibit systematic patterns regarding next-mention biases but do not affect pronoun production biases. This again broadens the empirical scope of the debate and evidence base for the strong form of the Bayesian Model.

2 Background

In this study, we contrast the Bayesian Model with two alternative models of pronoun interpretation: the Mirror Model, which posits that listeners interpret pronouns to refer to the referents that speakers choose to mention with pronominal referring expressions, and the Expectancy Model, according to which listeners' interpretation bias toward a referent is their estimate that the referent will get mentioned next. This section first introduces these three distinct models of pronoun interpretation. We also address the debate regarding whether pronoun production is influenced by the next-mention bias induced by meaning-driven factors. The *strong* form of the Bayesian Model predicts that pronoun production is insensitive to these factors. Following that, we provide an overview of the experimental paradigm and materials used in previous studies that are relevant to our research. We also provide a concise summary of the findings from studies that explore similar questions to ours. Finally, we outline the objectives of the present study.

2.1 Mirror Model

A common wisdom shared amongst discourse researchers is that interlocutors represent the ongoing discourse by constructing a mental model and continually updating it as they process the discourse (e.g., Lambrecht 1996). As discourse unfolds, representations of certain discourse referents are likely to be more active in memory and attention than others.

Traditional approaches to discourse anaphora posit that referents can be ranked according to the activation status of their mental representations in memory. This activation status is thought to guide speakers' choice of which referent to mention and the type of referring expression to use. These approaches propose hierarchies mapping different referential forms to various activation statuses of referents (e.g. Givón 1983; Ariel 1990; Gundel et al. 1993). In general, they tend to associate more reduced expressions such as pronouns with referents that are more activated in memory and attention i.e. more salient (see Table 1 for the Givenness Hierarchy in Gundel et al. 1993 as an example). The underlying assumption is that when referents are easily accessible in memory for both speaker and listener, facilitated by contextual and cognitive information, speakers can more effectively use less explicit referring expressions. As such, the choice of referring expression is driven by speakers' assumptions regarding the cognitive status of

in focus	**> activated >**	**familiar >**	**uniquely identifiable**	**> referential >**	**type identifiable**
it	this/that/this N	that N	the N	indefinite this N	a N

(Gundel et al. 1993)

Table 1: Givenness Hierarchy

the intended referent in listeners' minds. This implies that both speakers and listeners use the same cues to determine referent salience and rely on a shared concept of referent salience for production and interpretation. Therefore, during interpretation, it is assumed that listeners reverse-engineer the speaker's production process, interpreting pronouns by considering what entities the speaker would most likely refer to using a pronoun as opposed to a competing referential form.

Following previous research, we adopt the term **Mirror Model** to refer to these types of approaches. In these approaches, pronoun production and pronoun interpretation align on the same notion of referent salience, essentially mirroring each other. We define this model as per Equation I, as done in prior studies (e.g., Rohde & Kehler 2014; Bader & Portele 2019; Zhan et al. 2020; Patterson et al. 2022). Interpretation bias, denoted as $P(referent \mid pronoun)$, represents the probability of a specific referent being the intended reference for a given pronoun. On the other hand, production bias, denoted as $P(pronoun \mid referent)$, represents the probability of a pronoun being used to refer to a particular referent. The sum in the denominator is computed over all possible referents such that $P(pronoun \mid referent)$ is calculated for each candidate referent and those probabilities are summed. This summation ensures that the probabilities across all possible referents add up to 1, normalizing the probabilities. However, for the purpose of our discussion, we can disregard this denominator as it acts as a constant factor (i.e., it is the same for $P(referent \mid pronoun)$ for all referents). Therefore, in the Mirror Model, the interpretation bias towards a referent is directly proportional to the likelihood of the speaker using a pronoun to refer to that referent i.e., production bias.

$$\text{(I)} \qquad P(referent \mid pronoun) \leftarrow \frac{P(pronoun \mid referent)}{\sum_{referent \in referents} P(pronoun \mid referent)}$$

2.2 Bayesian Model

Challenging the assumption of a unified salience notion in pronoun production and interpretation, the proposal put forth by Kehler et al. (2008) and Kehler & Rohde (2013) suggests a Bayesian framing for the relationship between pronoun interpretation and production. This model provides a plausible explanation for the observed asymmetries between pronoun production and interpretation in empirical studies (e.g., Rohde & Kehler 2014; Mayol 2018). For example, story continuation data from Stevenson et al. (1994) found no strong interpretation bias for the ambiguous pronoun *he* in (1a) towards either the subject *John* or the non-subject *Bill* (a roughly 50/50 interpretation preference). However, for (1b), there was a strong bias towards using a pronoun when participants referred to the previous subject *John*, and a strong bias toward using a name when they referred to a non-subject *Bill*.

(1) a. John passed the comic to Bill. He __________
b. John passed the comic to Bill. __________

According to the Bayesian Model, this asymmetry can be characterized using Bayes' Rule, as shown in Eq. II. Such a formulation allows for the differentiation between the bias in pronoun production observed by Stevenson et al. (1994) and the pattern of pronoun interpretation. While the latter is related to the production bias, it also incorporates the next mention bias.

(II) $$P(referent \mid pronoun) = \frac{P(pronoun \mid referent)P(referent)}{\sum_{referent \in referents} P(pronoun \mid referent)P(referent)}$$

More specifically, pronoun interpretation bias, represented by $P(referent \mid pronoun)$ in the formulation, is determined by two probabilities: (i) $P(referent)$, the listener's estimate that a referent will get mentioned next, and (ii) $P(pronoun \mid referent)$, the listener's estimate that the speaker will use a pronoun to refer to that referent. In other words, the Bayesian Model predicts that if we have separate estimates of $P(referent)$, $P(pronoun \mid referent)$, and $P(referent \mid pronoun)$, then we would expect Eq. (II) to hold approximately. This claim is known as the weak claim of the Bayesian Model, henceforth referred to as **Weak Bayes** (e.g., Kehler & Rohde 2013; Rohde 2019; Zhan et al. 2020; Hoek et al. 2021; Patterson et al. 2022).

The strong form of the Bayesian Model, henceforth referred to as **Strong Bayes**, further specifies the types of contextual factors that affect each term on the right-hand side of Eq. II (see the references just given for Weak Bayes). Strong Bayes reflects an empirical observation that a set of meaning-driven factors (e.g., verb type, coherence relations) influence the next-mention bias and accordingly affect the pronoun interpretation bias. However, the speaker's decision regarding the pronominalization of a referent is insensitive to these factors. Instead, pronoun production is primarily influenced by grammatical factors (e.g., subjecthood) and/or factors associated with information structure (specifically, topichood). These factors together essentially lead to a preference for sentential subjects.[1]

However, empirical studies have produced conflicting evidence concerning whether certain semantic and pragmatic factors like verb type and discourse relation, actually influence pronoun production. In the following section, we will present the Expectancy Hypothesis, an alternative theory to Strong Bayes, along with a pronoun interpretation model derived from this theory, known as the Expectancy Model.

2.3 Expectancy Model

In contrast to Strong Bayes, the Expectancy Hypothesis (Arnold 1998; 2001; Arnold et al. 2007; Arnold 2010; Arnold & Tanenhaus 2011) posits that both next-mention bias and pronoun production bias are influenced by meaning-driven factors. According to the Expectancy Hypothesis, next-mention bias, which refers to the expectations about the next mention from listeners, is closely tied to speakers' choice of referring expression. Specifically, it suggests that listeners' estimate of the likelihood that a particular referent will be mentioned next strongly influences the activation level of that referent in the interlocutors' mental representation of discourse. Speakers can thus calculate the former as an estimate of the latter, using more reduced forms, such as pronouns, for referents that are expected or highly predictable to their listeners.

[1] Note that this study does not explicitly differentiate whether production biases stem from subjecthood or topichood. For further insights into this matter, see studies like Kehler & Rohde (2013), Rohde & Kehler (2014) and Zhan et al. (2020).

Furthermore, as the Expectancy Hypothesis suggests a direct link between listeners' estimate of the likelihood that a referent will be continued in the discourse and the activation level of that referent in their mental representation (referent accessibility), it predicts that reference interpretation is strongly influenced by the expectancy of a referent. Given this and following previous research, we define a third model of pronoun interpretation, termed as *Expectancy Model.* In this model, pronoun interpretation bias primarily depends on the next-mention bias, as shown in Eq. III.

$$P(referent \mid pronoun) \leftarrow \frac{P(referent)}{\sum\limits_{referent \in referents} P(referent)} \tag{III}$$

In the following section, we present the methods and materials employed in previous studies, which bear upon various aspects of the present study.

2.4 Methods used in previous studies

Previous empirical investigations on pronoun production and interpretation have primarily employed story continuation tasks with factorial designs and with semantic manipulations. In the standard experimental paradigm, participants are presented with a controlled context and asked to provide a natural continuation to it. As both comprehenders and producers, participants must first understand the context, such as Example (2a) and (3a) below, and then provide a continuation based on how they expect the story to proceed, as in the sample continuations in (2b) and (3b).

The prompt type is manipulated by alternating between bare forms and pronouns. In bare-prompt conditions, such as (2a), participants are expected to provide an entire sentence as a continuation e.g., (2b). These responses are then annotated by judges for the following two choices: (1) the choice of referent, specifically the character to which participants first make reference in their continuations (e.g., Amanda, Brittany, both or neither); (2) the choice of referential form, which concerns whether the expression participants used for their first reference is a pronoun, a name or a more explicit description. Data from the bare-prompt conditions allow for the calculation of probability estimates for the three models. Specifically, the next-mention bias, $P(referent)$, can be computed from the annotations of referent choice, and the pronoun production bias, $P(pronoun \mid referent)$, can be computed when considering annotations of both referential form and referent.

On the other hand, in pronoun-prompt conditions, such as (3a), participants are presented with an ambiguous pronoun at the beginning of a sentence and are then tasked with completing the rest of the sentence, such as (3b). In these instances, responses are annotated solely for the choice of referent, identifying the character that the participants interpret the given pronoun to represent. This provides actual pronoun interpretation biases. Hence, the predicted interpretation biases (derived from bare-prompt data) can be compared against the observed interpretation biases (derived from pronoun-prompt data).

(2) a. Amanda amazed Brittany. ________ [Bare prompt]
b. She was not expecting that kind of performance.

(3) a. Amanda amazed Brittany. She ________ [Pronoun prompt]
b. was always a good dancer.

In previous studies that test coreference biases, the experimental stimuli were mainly created by manipulating two factors: verb types and discourse relations. In addition to offering measurements of next-mention rates, pronoun production rates, and interpretation rates, which can be utilized to evaluate models of pronoun interpretation, these factors were chosen because they provide contexts with opposing next-mention biases to one referent or the other. This allows researchers to examine the influence of these factors on pronoun production biases and test the prediction of Strong Bayes and the Expectancy Hypothesis. We discuss these factors below.

2.4.1 Verb semantics

The sensitivity of next-mention coreference biases to verb semantics has been widely investigated (e.g., Arnold 2001; Fukumura & Van Gompel 2010; Rohde & Kehler 2014; Mayol 2018; Zerkle & Arnold 2019; Weatherford & Arnold 2021).

One verb type that has been used is transfer-of-possession verbs. For instance, the main verb *give* in the sentence *Lisa gave the leftover pie to Brendan* expresses a transfer event and assigns thematic roles of Source and Goal to participants in the event. The Source role identifies the object from which motion of transfer proceeds, while the Goal identifies the object towards which transfer proceeds (Stevenson et al. 1994). These verbs can be divided into two subgroups with symmetric argument structures: Source-Goal verbs such as *give* in (4), and Goal-Source verbs like *catch* in (5).

(4) Lisa$_{source}$ **gave** the leftover pie to Brendan$_{goal}$. ________

(5) Lisa$_{goal}$ **caught** a cold from Brendan$_{source}$ two days before Christmas. _____

Several story continuation experiments have shown a consistent tendency for participants to refer back to the Goal referent more frequently than to the Source referent, across both types of verbs (e.g., Stevenson et al. 1994; Arnold 2001; Rosa & Arnold 2017). According to Stevenson et al. (1994), this next-mention bias stems from a natural focus on the consequences elicited by verbs that semantically depict transfer events. For instance, in (4), participants tend to talk about what the Goal referent *Brendan* did next after receiving *leftover pie*. To test how this next-mention bias induced by verb semantics influences pronoun production while controlling for the well-known effects of grammatical roles, previous studies compare the pronominalization of the Goal and Source when both are introduced in the same grammatical position (e.g., Lisa in (4) vs. Lisa in (5), which are both mentioned in subject position).

Implicit causality verbs, such as *impress* or *admire*, are another well-tested and frequently used verb type in manipulation. These verbs describe a mental state and assign two thematic roles: a Stimulus, which is the argument that gives rise to the psychological state, and an Experiencer, which is the argument that experiences the psychological state. Like transfer-of-possession verbs, these verbs present crossed argument structures (see examples (6)-(7)), and they also elicit strong next-mention biases, but this time towards the Stimulus.

(6) David$_{stimulus}$ **impressed** Linda$_{experiencer}$. ________

(7) David$_{experiencer}$ **admired** Linda$_{stimulus}$. ________

Studies have shown that these verbs induce biases towards the Stimulus role when providing an explanation for the cause of an event, which might fall on either the subject or object position (e.g., Stevenson et al. 1994; Fukumura & Van Gompel 2010; Ferstl et al.

2011; Rohde & Kehler 2014; Mayol 2018; Zhan et al. 2020). For example, in the case of *impress* in (6) there is a strong preference for referring back to the Stimulus, David, who is in the subject position, while for *admire* in (7), continuations also preferably refer to the Stimulus, Linda, who is mentioned in the object position.

As in transfer-of-possession contexts, the bias towards the Stimulus over the Experiencer in implicit causality scenarios is used to test whether bias induced by verb semantics affects pronoun production e.g., are there more pronouns produced referring to the Stimulus subject referent David in (6) than to the Experiencer subject referent David in (7)?

2.4.2 Discourse relations

Researchers have used discourse relations as another factor to manipulate next-mention biases in addition to verb semantics. Discourse relations hold between clauses and can be implicitly inferred or explicitly marked by connectives. For example, the statement *John left* can be connected to *Mary stayed* by Explanation (*John left (because) Mary stayed*) or Result (*John left (so) Mary stayed*) or Contrast (*John left (but) Mary stayed*). By manipulating the connectives, previous research has shown that their semantics can interact with verb semantics to shape the preference for the upcoming referent (e.g., Fukumura & Van Gompel 2010; Holler & Suckow 2016; Hwang et al. 2022). For instance, while, as we have seen, speakers tend to continue segment (7) with the Stimulus, Linda, in an explanation (*because*…), they tend to continue it with the Experiencer, David, when talking about the result of the event (*so*…; see example (8)). The latter is opposite to the default next-mention bias elicited by implicit causality verbs, which is towards the Stimulus. Similar sensitivity to discourse relations has been reported for transfer-of-possession verbs (Stevenson et al. 1994; Kehler et al. 2008).

(8) David$_{experiencer}$ admired Linda$_{stimulus}$ **so** ________

While some studies focus on how discourse relations modulate transfer-of-possession and implicit causality biases (e.g., Stevenson et al. 1994; Fukumura & Van Gompel 2010; Holler & Suckow 2016; Hwang et al. 2022; Hwang 2023a), Hwang (2023b) extends the investigation of how discourse relations affect next-mention expectations to contexts beyond transfer-of-possession and implicit causality. Hwang (2023b) explores the role of connectives in facilitating a general sense of subject continuity and action continuity (see also Kehler 2002). They found that connectives of Narration, such as *and (then)*, better support this continuity than connectives of other relations, such as *while*. In other words, using *and (then)* to link clauses creates a stronger expectation for a continuation of the same subject and action than using other types of connectives (see (9) for a Korean example from Hwang 2023b). While Hwang (2023b) examined the production of zero pronouns in discourse relations marked by distinct connectives, the potential impact of discourse relations on overt pronoun production remains underexplored.[2]

[2] Note that both in Hwang (2023b) and in the current study, as further discussed in Section 3, the subject referents are more predictable in contexts describing a sequence of actions or events, while their predictability decreases in contexts with lower action continuity. In contrast, a recent study by Hwang & Lam (2023) examines scenarios where subject referents are less predictable in contexts with higher action continuity, thereby disentangling predictability from action continuity.

(9) **Minswu-ka Hyenwu-wa palphyo cwunpi-lul ha-ko/nuntey** ________.
Minsu-NOM Hyunwoo-with presentation preparation-ACC do-and/while ________
'Minsu prepared a presentation with Hyunwoo and (then)'/ 'While Minsu was preparing a presentation with Hyunwoo' ________

2.5 Previous findings

This section summarizes findings from previous studies that aimed to quantitatively assess the three models of pronoun interpretation and evaluate the Strong Bayes hypothesis. These studies employed standard experimental paradigms with bare and pronoun prompts (e.g., Ex. (2) and (3)). Their findings are directly relevant to our research questions, which we will elaborate on in Section 3.

The study by Rohde & Kehler (2014) marked the first quantitative evaluation of the three pronoun interpretation models. Their approach involved manipulating verb types by using subject-biased and object-biased implicit causality verbs in English (e.g., Ex. (6) and (7)). They assessed the correlations between predicted interpretation biases (obtained from bare-prompt data) and observed biases (collected from pronoun-prompt data) using R^2. Further details on this metric, along with two others below, are provided in Section 5.4. Notably, the Bayes-derived estimates consistently exhibited the strongest correlation with observed data. Additionally, they observed the influence of verb type on next-mention and pronoun interpretation biases, while pronoun production biases remained unaffected, giving rise to the Strong Bayes hypothesis.

Expanding the research scope, Zhan et al. (2020) investigated overt and null pronouns in Mandarin Chinese. They conducted three experiments using implicit causality verbs and introduced additional variables such as voice (active vs. passive) and syntactic construction in the second and third experiments. No evidence supporting the influence of verb type on pronoun production biases was found throughout the study, consistent with the Strong Bayes hypothesis. To evaluate pronoun interpretation model performance, they employed R^2 alongside Mean Squared Error (MSE) and Average Cross Entropy (ACE) metrics, offering a more comprehensive assessment. Across nine evaluations, the Bayesian model outperformed both the Mirror Model and the Expectancy Model, ranking best in six and second best in three.

In a more recent study, Patterson et al. (2022) further demonstrated the superior predictive accuracy of the Bayesian Model for the German personal pronoun *er* and the German demonstrative pronoun *dieser* when compared to the Mirror Model and the Expectancy Model. They improved the assessment of model performance by incorporating Bayesian methods that propagated data uncertainty to predictions, producing distributions of possible values rather than point estimates like the ones in the metrics employed by Zhan et al. (2020) and Rohde & Kehler (2014). Furthermore, Patterson et al. evaluated the Strong Bayes hypothesis through contrasts between accusative and dative verbs, as well as stimulus-experiencer versus experiencer-stimulus verbs. In both cases, no interaction between verb type and referent was found in pronoun production likelihoods, supporting the Strong Bayes hypothesis.

In summary, the body of evidence consistently points towards the Bayesian Model outperforming the other two models across various languages and experimental setups. Furthermore, these studies also consistently provide support for the Strong Bayes hypothesis.

3 Goals and hypotheses of the current study

Story continuation experiments, as previously discussed, have been employed in prior empirical research (Rohde & Kehler 2014; Bader & Portele 2019; Zhan et al. 2020; Patterson et al. 2022) to evaluate the Bayesian Model. These experiments employed targeted materials featuring specific verb types as stimuli. In the present study, our objective is to assess the performance of the Bayesian model, both its weak and strong forms, in more naturalistic contexts, using naturally occurring language. Our investigation is guided by two primary research questions.

One question pertains to evaluating which model for pronouns (i.e., Bayesian, Expectancy, or Mirror) best accounts for the interpretation bias observed for ambiguous pronouns. We explore this question in a passage continuation experiment (Section 5). To achieve this, we compare the predictions of these three proposed models of pronoun interpretation against the observed interpretation biases in continuations elicited in passages extracted from naturally occurring passages.

Another question addresses the prediction of Strong Bayes. We investigate this using both observational and experimental data (sections 4 and 5). Specifically, we examine whether pronoun production biases remain unaffected by meaning-driven factors —here, discourse relations —that have been shown to influence biases for next mention, as predicted by Strong Bayes. Our decision to focus on discourse relations in the analysis is driven by our corpus-based methodology. The rationale behind this choice will be further elaborated upon in Section 4.

We identified three discourse relations for which we have specific hypotheses regarding their differing next-mention biases, which in turn allows for testing whether pronoun production is similarly influenced by these discourse relations. The selection of discourse relations for analysis was guided by the classification proposed by Kehler (2019), which is adapted from the work of Hobbs (1990). The three relations we concentrate on are Narration, Contrast, and Result.[3] Examples illustrating these three relations can be found in Table 2.

Discourse relation	Example
Narration	Judas went over to Jesus. Then he kissed him.
Result	Hurricane Maria struck Puerto Rico. As a result, the country is facing a desperate humanitarian crisis.
Contrast	The Constitution does not expressly give the president such power. However, the president does have a duty not to violate the Constitution.

Table 2: Examples for the discourse relations of interest, sourced from the OntoNotes corpus.

We excluded relations that are typically signaled by ambiguous connectives, such as Parallel. Parallel, which signifies the presence of discourse segments that share similar or parallel content, structure, or form, is typically marked by *and* such as in *Set stack A empty and set link variable P to T*, but this connective is compatible with a wide range of other relations. In addition, we also excluded relations that are most frequent in subordinating

[3] The term *Narration* is used in Segmented Discourse Representation Theory (Asher & Lascarides 2003); Kehler (2002) calls it *Occasion*, and Rhetorical Structure Theory uses the term *Sequence* (Mann & Thompson 1988).

constructions, such as Explanation (Asher & Vieu 2005), as these constructions generally lead to higher pronoun production across the board (Fukumura & Van Gompel 2010).

For our research question regarding the prediction of Strong Bayes, we put forward the following hypotheses.

Strong Bayes Hypothesis 1: Next mention bias. Concerning the differing next-mention biases induced by the three discourse relations, our hypothesis is that we will find a higher percentage of *subject* instances in Narration than in the other two relations; that is, we expect a greater likelihood of the subject being re-mentioned in Narration than in Contrast and Result. The justification for this hypothesis is rooted in the inherent characteristic of the Narration relation that maintains continuity in the entities, typically the topical subject, around which narrative sequences of events are constructed. In English, the grammatical subject position serves as the traditional locus for introducing the topic (Gundel 1988; Lambrecht 1996; Ariel 2001). In contrast, this propensity is less pronounced in other relations; specifically, in a Result-oriented discourse like *Hurricane Maria struck Puerto Rico yesterday. As a result* ..., it is quite plausible that the narrative will next turn to the patient role *Puerto Rico* that bears the brunt of the circumstances. This exploration of subject continuity in narrative discourse parallels the work of Hwang (2023b), which demonstrated that the Korean connective *-ko* 'and (then)' tends to maintain stronger subject continuity, compared to the connective *-nuntey* 'while'. In this study, we focus on English contexts and extend the investigation by comparing *and (then)* and other Narration connectives with a set of connectives that signal Result and Contrast. Note that we expect a larger percentage of references to the previous subject than to non-subjects across *all* relations, due to the strong effect of subjecthood on reference continuation (e.g. Arnold 2001). Yet, we predict that this effect to be particularly prominent in the case of Narration.

Strong Bayes Hypothesis 2: Pronoun production bias. This hypothesis can be tested only if Hypothesis 1 is supported by the data. Strong Bayes predicts uniform pronoun production rates for referring to the subject across the three relations —Narration, Contrast, and Result —despite their differing propensities for subject continuation. Alternatively, according to the Expectancy Hypothesis, discourse relations are predicted to influence not only next-mention bias but also pronoun production bias, leading to a higher pronominalization rate for subject re-mentions in the Narration relation than in the other two relations.

In the following sections, we embark on an initial exploration of these questions via corpus-based analyses. Our preliminary findings demonstrate alignment with the prediction of the Strong Bayes model, which posits that factors like discourse relations that can affect next-mention bias (Hypothesis 1) do not affect pronoun production bias (Hypothesis 2).

To obtain a more robust set of evidence supporting our initial findings on Strong Bayes, as well as to obtain interpretation data for evaluating Weak Bayes, we used a set of corpus passages as stimuli in a story completion experiment with human participants. The bare-prompt data derived from this experiment provide the basis for examining the relationship between next-mention bias and pronoun production bias. Later, the same bare-prompt data are used to generate predicted interpretation rates for the Bayesian, Mirror, and Expectancy models. We evaluate these models by comparing their predictions to the observations from the pronoun-prompt conditions.

4 Observational examination of Strong Bayes: corpus analyses

To test Strong Bayes and Hypotheses 1 and 2, we automatically retrieve naturally produced texts from two corpora annotated in computational linguistic research. The objective is to examine whether pronoun production in these corpus texts exhibits sensitivity towards meaning-driven factors that are shown to influence next-mention expectations. We specifically test the prediction of Strong Bayes, which posits that such sensitivity is absent.

We operationalize next-mention expectations in terms of re-mention frequency across corpus contexts. This approach is based on the assumption that hearers track statistical regularities in their input in order to predict upcoming information. Corpus data, in this regard, is considered to capture the distributional patterns that have been used to give estimates of expectations or predictability regarding upcoming information (Frank et al. 2013; Verhagen et al. 2018; Guan & Arnold 2021). Under this premise, a referent (such as the grammatical subject) that is re-mentioned more frequently in a class of corpus contexts compared to other classes is deemed more predictable for comprehenders within that kind of context.

Initially, our approach involved the extraction of contexts that closely resembled the stimuli used in previous studies for story continuation tasks: sentences featuring a transfer-of-possession verb as the main verb, along with a goal referent, a source referent, and a theme referent; and sentences containing an implicit causality verb and its corresponding arguments. However, differences in corpus texts compared to the controlled stimuli used in psycholinguistic experiments made it not possible for us to use this kind of semantico-pragmatic context. The details of the analyses on verb types and further explanation regarding the difficulties encountered can be found in Section A of the Supplementary file.

Therefore, our focus shifted toward investigating discourse relations. Previous research has shown that next-mention biases for verbs are modulated in interaction with different discourse relations. We extend this work by testing discourse relations across the board (see the recent study on Korean by Hwang 2023b), with the hypothesis that expectations primarily driven by discourse relations might be more robust in corpora than those induced by specific verb types. This is because while effects with specific verbs are attested only in very strict contextual conditions, as discussed in the Supplementary file, discourse relations can be expected to have a similar semantic general effect across contexts.

We formulate hypotheses in terms of the differing next-mention biases across three discourse relations –Narration, Contrast and Result, as explained in Section 3. To provide a comprehensive picture of both discourse relations signaled by discourse connectives and those manually identified by human annotators, we conducted analyses on two separate corpora. The first corpus contains rich linguistic information but lacks explicit discourse structure annotations; hence, we resorted to using explicit discourse connectives as signals to extract passages featuring specific discourse relations. We carried out a second analysis with a smaller corpus that features manual annotations of discourse relations, irrespective of the presence of an explicit connective. This allows us to take into account both explicit and implicit connections between clauses.

4.1 Corpora

4.1.1 OntoNotes

We first draw upon OntoNotes (Weischedel et al. 2013), a corpus that is widely used in Computational Linguistics for research on the computational modeling of anaphora. We restricted our focus to the English segment of OntoNotes, which consists of approximately 1.7 million words encompassing data from a variety of genres, as detailed in Table 3.

Genre	Size
Newswire	625K
Broadcast news	200K
Broadcast conversations	200K
Web data	300K
Telephone conversation	120K
New Testament and Old Testament	300K
Total	1745K

Table 3: English portion of OntoNotes: genres and corresponding sizes.

OntoNotes comes with rich manual annotations that enable its use for the purposes of the present article, exemplified in Table 4. In particular, we leveraged annotations related to coreference (anaphoric relations) and morphosyntactic information to automatically extract contexts of the discourse relations of interest. More specifically, the coreference chains, visualized in Example (10), enable the automatic identification of mentions that refer to the same entity, thereby facilitating the estimation of re-mention frequency. The syntactic parse trees allow for the identification of the grammatical roles of mentions, distinguishing between grammatical subjects and non-subjects.

(10) **[A wildfire in California]** $_0$ forced **[hundreds of people]** $_1$ from **[their]** $_1$ homes.

Word	POS	Tree	Lemma	Them. role	Sense	Speaker	Named entity	predicate-argument	Coreference
A	DT	(TOP(S(NP(NP*	-	-	-	-	*	(ARG0*	(0
wildfire	NN	*)	-	-	-	-	*	*	-
in	IN	(PP*	-	-	-	-	*	*	-
California	NNP	NP*)))	-	-	-	-	(GPE)	*)	0)
forced	VBD	(VP*	force	01	1	-	*	(V*)	-
hundreds	NNS	(NP(NP*)	-	-	-	-	(CARDINAL)	(ARG1*	(1
of	IN	(PP*	-	-	-	-	*	*	-
people	NNS	(NP*)))	people	-	1	-	*	*)	1)
from	IN	(PP*	-	-	-	-	*	(ARG2*	-
their	PRP$	(NP*	-	-	-	-	*	*	(1)
homes	NNS	*)))	home	-	1	-	*	*)	-

Table 4: Multiple layers of annotation in OntoNotes.

4.1.2 Rhetorical Structure Theory Discourse Treebank (RST-DT)

RST-DT consists of 385 Wall Street Journal articles (176k words) from the Penn Treebank (Marcus et al. 1993), annotated with discourse relations in the framework of Rhetorical

Structure Theory (RST, Mann & Thompson 1988).[4] Under the RST framework, texts are represented as a tree and are broken down into minimal discourse units (often corresponding to clauses), which are called elementary discourse units (EDUs). As illustrated in Figure 1, each leaf of the tree corresponds to an EDU. Adjacent EDUs are connected by discourse relations to form larger segments.

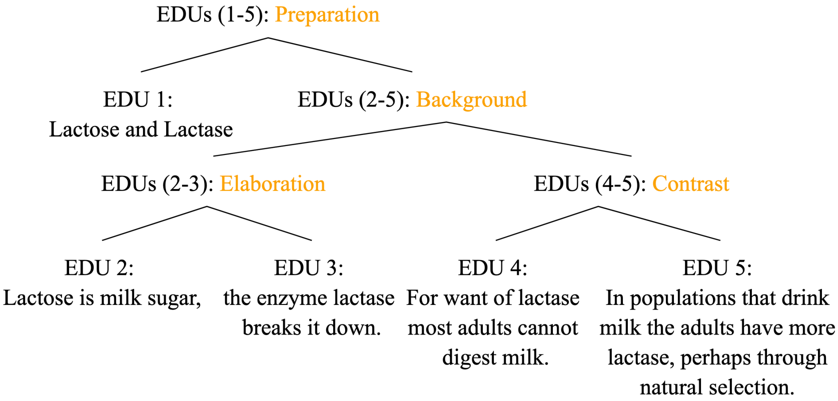

EDUs (1-5): Preparation
EDU 1: Lactose and Lactase
EDUs (2-5): Background
EDUs (2-3): Elaboration
EDUs (4-5): Contrast
EDU 2: Lactose is milk sugar,
EDU 3: the enzyme lactase breaks it down.
EDU 4: For want of lactase most adults cannot digest milk.
EDU 5: In populations that drink milk the adults have more lactase, perhaps through natural selection.

Figure 1: Graphical representation of an RST analysis (own production using text from the RST website, www.sfu.ca/rst).

The inventory of discourse relations annotated in RST-DT is fairly fine-grained, with 78 relations in total. In our study, we use a coarser-grained taxonomy of relations (Carlson & Marcu 2001), in which the 78 RST relations are partitioned into 16 broad categories based on their rhetorical similarity. For instance, one of the major categories is *Contrast*, which is the umbrella term for the relations *Contrast*, *Concession*, and *Antithesis* in the original inventory.

4.2 Method

4.2.1 Extraction of explicitly signaled relations from OntoNotes

To extract passages of discourse relations from the OntoNotes corpus, we relied on explicit connectives, given the absence of discourse structure annotations. The connectives used for our passage extraction are listed in Table 5. Our selection of connectives was guided by the distribution patterns of both explicit and implicit connectives inserted by human annotators reported in the Penn Discourse Treebank 3.0 Annotation Manual (Webber et al. 2019).[5] These selected connectives primarily signal the target relation rather than other relations, thus rendering them mostly non-ambiguous.[6]

[4] Out of these 385 Wall Street Journal articles in RST-DT, 277 are also present in the OntoNotes corpus. However, we use the manual annotation, which includes implicit and explicit relations. This analysis thus presents evidence that is complementary to the previous one with OntoNotes.

[5] The connectives we have selected correspond to the following relations annotated in Penn Discourse Treebank 3: Result → Contingency.Cause.Result; Contrast → Comparison.Contrast, Comparison.Concession.Arg2-as-denier; Narration → Temporal.Asynchronous.Precedence, as these most closely align with the targeted relations in our study.

[6] The connective *so* often presents polysemy, which introduces some degree of ambiguity. However, we included it because it is very commonly used to indicate the Result relation. In order to attenuate potential inconsistencies from the diverse semantics of *so*, we have restricted its part of speech to an adverb (part-of-speech tag *RB* in OntoNotes), rather than being tagged as a preposition or subordinating conjunction (tag *IN*). Additionally, we consider the surrounding tokens within the extracted passages, such as excluding instances of *so far*.

Discourse relation	Connectives
Narration	afterward, afterwards, later, next, (a period of time) later/after, after it/that, subsequently, (and) then, thereafter
Result	(and) so, thus, accordingly, consequently, hence, therefore, as a result, as a consequence
Contrast	in contrast, in comparison, but, yet, by comparison, by contrast, conversely, however, nevertheless, nonetheless, on the contrary, on the other hand

Table 5: Connectives used for passage extraction.

We focused on cases of sentence-initial coordinating conjunction, in which the connective appeared at the beginning of a sentence, such as *Judas ate the bread Jesus gave him. **Then** he immediately went out.* We left out intra-sentential cases such as *An evil spirit comes into him, **and then** he shouts*. This decision was motivated by the fact that sentence-internal coordinating conjunctions often co-occur with null subjects (or verb phrase coordination constructions), as in *Judas went over to Jesus and then Ø kissed him*. Exploring this phenomenon is beyond the scope of our current study and could be an avenue for potential future research.

After identifying discourse relations using connectives, we identified the next mention, defined as the matrix clause subject right after the connective.[7] The extracted contexts were then automatically classified into three types: *subject*, when the next mention coreferred with the preceding subject, *non-subject*, when it coreferred with another element in the preceding clause, and *other* when it coreferred with referents that have not been mentioned in the preceding clause (either new referents, or referents from earlier discourse). Table 6 lists examples for the Result relation. Note that, unlike typical psycholinguistic experiments in this field, and like previous work using corpora (Arnold 2001; Guan & Arnold 2021), we consider references to entities that are not in the previous clause (*other*). This decision is motivated by the fact that OntoNotes offers much richer contexts compared to controlled stimuli, increasing the likelihood of re-mentioning referents beyond those in the previous clause. These cases should be included as comparison points.

4.2.2 Extraction of manually-annotated relations from Rhetorical Structure Theory Discourse Treebank

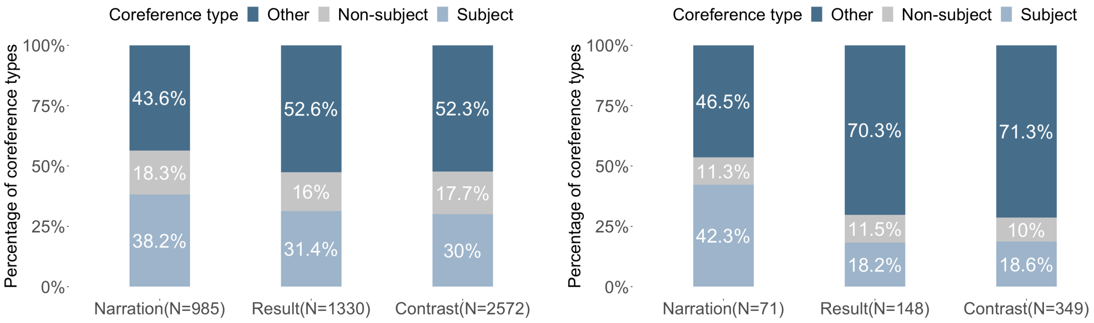
The RST-DT corpus provides annotations for relations at both the intra-sentential level, involving small segments within sentences, and the inter-sentential level, encompassing larger segments across sentences. To maintain comparability with our previous analysis using the OntoNotes corpus as well as psycholinguistic experiments, we specifically focus on inter-sentential samples. These samples consist of relations where the left-hand argument and the right-hand argument are adjacent, but the right-hand argument begins as a separate sentence.

We selected RST relations that align approximately with the relations we extracted in OntoNotes, namely Narration, Contrast, and Result. The mapping between the origi-

[7] Using the first noun phrase after the connective instead results in too much noise, such as cases in which it indicates time or location (e.g. *this week* or *school* in *at school*).

Coreference type	Example
subject	Winning candidate Chen Shuibian captured only 39% of the vote. As a result, he must take a moderate line that stresses inter-party cooperation.
non-subject	Then Zechariah could not speak to them. So the people knew that he had seen a vision inside the Temple.
other	The navy of Iraq has a terrific commander. So the people around him they'll follow him into battle.

Table 6: Automatic labeling paradigm of coreference types, exemplified with the Result relation.



(a) OntoNotes

(b) RST-DT

Figure 2: Coreference type by discourse relation in OntoNotes (left; total samples = 4,887) and RST-DT (right; total samples = 568).

nal taxonomy of RST-DT and our categorization is provided in Table **??**, which can be found in Section B.3 of the the Supplementary file. For clarity, we maintain the same terminology as in the OntoNotes analysis when presenting our findings.

The data extraction process was again conducted entirely automatically. It is important to note that while RST-DT does not include coreference annotations, the Anaphora Resolution and Underspecification corpus (ARRAU; Poesio et al. 2013) provides coreference annotations for the same set of articles as those in RST-DT. Therefore, we aligned these two corpora to extract both discourse relations and coreference data.

As in the previous analysis, the extracted contexts for each relation were categorized into three groups: *subject coreference*, *non-subject coreference*, and *other coreference*. For more details on our extraction strategy, see Section B.4 of the Supplementary file.

4.3 Results

The results from both corpora support Hypothesis 1: the subject referent is more frequently re-mentioned in Narration than in other relations, as follows. Figure 2 shows the distribution of coreference types by relation in OntoNotes and in RST-DT, where *subject coreference* is higher in Narration than in Result and Contrast (raw counts for each type are presented in Section B of the Supplementary file).

We built mixed-effects logistic regressions where the dependent measure was whether the context continued with the subject referent or not, and a fixed effect for the 3-level discourse relation type, with Contrast as the reference level.[8] In our analysis of the OntoNotes sample, random intercepts for individual verbs were included, along with random slopes for relation types by verb, given that the effect of discourse relation types have been shown to vary across different verbs (e.g., Kehler et al. 2008). The analysis of the RST-DT sample, on the other hand, included only random intercepts for verbs. Random slopes for relations by verb were not incorporated due to insufficient data to accurately estimate them. In both analyses, we also included random intercepts for the document ID to account for potential variations associated with e.g. author style.

The results of the statistical analysis are reported in Table 7 and 8 for OntoNotes and RST-DT respectively. The effect of Narration on the likelihood of subject re-mention in both analyses is positive and significant at the chosen .05 alpha level, indicating that the subject referent is more frequently mentioned again in the Narration relation compared to the Contrast relation. Pairwise comparisons using the estimated marginal means show the subject referent is more frequently re-mentioned in Narration compared to Result (OntoNotes: $\beta = 0.50$, z $= 3.14$, $p = 0.005$; RST-DT: $\beta = 1.37$, z $= 3.77$, $p < 0.001$), while there is no difference between Contrast and Result (OntoNotes: $\beta = 0.13$, z $= 1.03$, $p = 0.56$; RST-DT: $\beta = -0.01$, z $= -0.02$, $p \approx 1$).

Effects		Estimate	SE	z	*p*
Intercept		−0.84	0.06	−13.62	
discourse relation	**Narration**	**0.37**	**0.12**	**3.16**	**0.002**
	Result	−0.13	0.13	−1.03	0.30

Table 7: Subject re-mention in OntoNotes: mixed-effects logistic regression model with the subject being re-mentioned as the dependent measure.

Effects		Estimate	SE	*z*	*p*
Intercept		−1.76	0.23	−7.69	
discourse relation	**Narration**	**1.37**	**0.33**	**4.22**	**<0.001**
	Result	0.01	0.27	0.02	0.98

Table 8: Subject re-mention in RST-DT: mixed-effects logistic regression models with the subject being re-mentioned as the dependent measure.

In terms of Hypothesis 2, we find no evidence that discourse relations affect the choice of referring expression, that is we find support for Strong Bayes, as follows. Figure 3 shows the raw data, which do not indicate a higher pronominalization rate in the Narration relation, despite the higher rate of subject re-mention in this relation; and this is supported by a statistical analysis. We conducted mixed-effects logistic regression analyses on *subject* and *non-subject* samples. The dependent measure in this analysis was whether the next mention was pronominalized or not. The fixed effects included discourse relation types, coreference types (subject or non-subject), and their interaction, with Contrast as the reference level for relation type and non-subject as the reference level for coreference type. In our analysis of the OntoNotes sample, we included random intercepts for both verbs and document IDs, consistent with the first set of models. Random slopes for relations by verb were not incorporated due to singular fit issues. In contrast, the analysis of the

[8] We used the glmer function from the lme4 package (v1.1-28; Bates et al. 2015) in R (R Core Team 2021).

RST-DT sample involved only random intercepts for document IDs, omitting random effects for verbs also due to singular fit concerns. As shown in Table 9 and 10, the analyses on both OntoNotes and RST-DT revealed no significant difference in pronominalization patterns among the three examined relations. Additionally, our findings replicate the widely attested observation that subjecthood affects pronominalization. Specifically, we observe a higher frequency of pronoun usage when referencing the preceding subject (see the three left bars in figures (a) and (b) in Figure 3) compared to non-subject entities (represented by the three right bars). We conducted an additional robustness test to check a potential confound related to analyzing pronoun production in corpus passages: whether the antecedent is a pronoun or not. We found that the rates of pronoun production do not exhibit variations across discourse relations, even after accounting for the influence of the antecedent's form (see more details in Section B.2 of the Supplementary file.

However, traditional null hypothesis testing aimed at identifying statistically significant differences, and this approach could be problematic when addressing hypotheses like Strong Bayes, which focus on the absence of an effect. This is because this method primarily assesses the presence rather than the absence of effects, potentially conflating true absence with undetected effects.

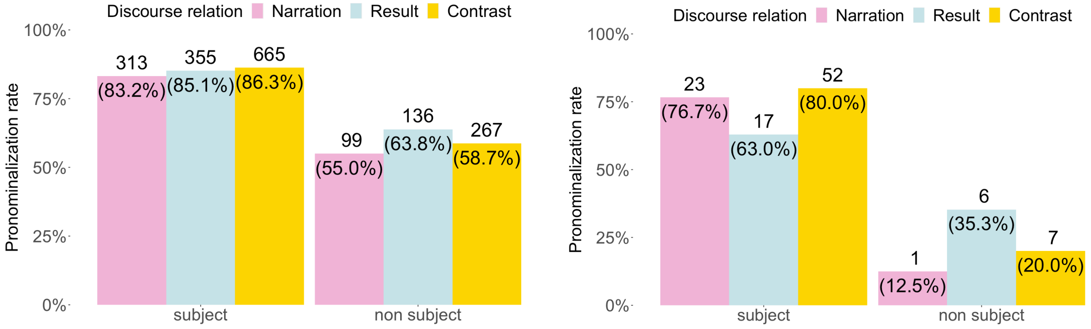
To address this, we adopted a Bayesian framework using Bayes Factors, which shifts the focus from finding evidence for a difference (i.e., rejecting the null hypothesis that there is no difference, as in traditional null hypothesis testing) to quantify support for true null hypotheses (e.g., Kass & Raftery 1995; Schad et al. 2022). Specifically, we assess how much more likely the data is under the model which represents the null hypothesis compared to the other which represents the alternative.

For both analyses with OntoNotes and RST-DT, we compared two Bayesian models: Model 1 (H1: alternative hypothesis) and Model 2 (H0: null hypothesis). Model 1 incorporates the same fixed effects and random effects as the models presented in Table 9 (for OntoNotes) and Table 10 (for RST-DT). Specifically, Model 1 incorporates both *Relation Type* and *Coreference Type* as predictors, as well as their interaction. Model 2, in contrast, excluded *Relation Type* and considered only *Coreference Type* as a predictor.

The models were fit using the brm function from the *brms* (v2.18.0; Bürkner 2017) package in R (R Core Team 2021). To ensure stable inferences, we utilized weakly informative priors, specifically Cauchy distributions with a center of 0 and a scale of 2.5 (Gelman et al. 2013). The fits were run with 4 chains, each comprising 4000 iterations, with half as warm-up. Prior to analysis, thorough diagnostic checks were conducted to rule out any potential pathologies in the estimation process.[9] For Bayes factors analysis, we used the *bridgesampling* package (Gronau et al. 2020; version 1.1.2) in R.

The results from both analyses strongly supported Model 2 (H0). In the analysis with OntoNotes, the Bayes Factor in favor of Model 2 (H0) over Model 1 (H1) was estimated to be 32,984 (in favor of Model 1 over Model 2: 0.00003). Similarly, in the analysis with RST-DT, the Bayes Factor in favor of Model 2 (H0) over Model 1 (H1) was estimated to be 86.75 (in favor of Model 1 over Model 2: 0.01). Therefore, our analyses indicate that the inclusion of *Relation Type* in the model does not contribute significantly to explaining pronoun production biases, and the simpler Model 2, representing the null hypothesis or Strong Bayes, is the preferable model in both cases.

[9] We verify that there are no divergent transitions; that all the $\hat{R}$ (the between- to within-chain variances) are close to one; that they had no saturated trajectory lengths (i.e., the sampler did not stop prematurely); that the number of effective sample size are at least 10% of the number of post-warmup samples.



(a) OntoNotes

(b) RST-DT

Figure 3: Pronominalization rate of next mention by relation in OntoNotes (left) and RST-DT (right). The graphs show the pronominalization rates when the next-mention corefers with the subject and when it corefers with a non-subject. Raw counts of samples with a pronominal next mention are presented on top of each bar, with the corresponding percentage in parentheses.

Effects		Estimate	SE	z	*p*
Intercept		0.23	0.14	1.69	
discourse relation	Narration	-0.04	0.21	-0.17	0.87
	Result	0.23	0.20	1.15	0.25
coreference	**subject**	**1.71**	**0.17**	**9.98**	**<0.001**
Narration:subject		-0.002	0.29	-0.007	0.99
Contrast:subject		-0.42	0.28	-1.46	0.14

Table 9: Pronominalization in OntoNotes: Mixed-effects logistic regression model with the next mention being a pronoun as the dependent measure.

Effects		Estimate	SE	z	*p*
Intercept		-1.73	0.56	-3.06	
discourse relation	Narration	-0.48	1.26	-0.39	0.70
	Result	0.85	0.75	1.13	0.26
coreference	**subject**	**3.22**	**0.73**	**4.44**	**< 0.001**
Narration:subject		0.36	1.39	0.26	0.79
Result:subject		-1.76	0.95	-1.85	0.06

Table 10: Pronominalization in RST-DT: Mixed-effects logistic regression model with the next mention being a pronoun as the dependent measure.

4.4 Interim discussion

To sum up, our corpus-based evidence confirms that discourse relations can induce varying degrees of bias in subject re-mentioning. Specifically, Narration exhibits a greater bias towards the continuation with a subject referent compared to Contrast and Result; and we do not find evidence that re-mention likelihood influences pronoun production within contexts of discourse relations. This disconnect between the likelihood of next mention and the likelihood of pronominalization aligns with the Strong Bayes theory.

As a side note, we observe from a comparison between the two figures in Figure 2 that, while the distribution of coreference types is comparable in the two corpora, there is a peculiar shortage of Narration contexts in the RST-DT corpus (71 samples only, which represents merely 12.5% of the total samples across all three relations, compared to 985 in OntoNotes, accounting for 20.2% of the total). This discrepancy is not likely to be due only to the difference in corpus size: OntoNotes is approximately ten times the size of RST-DT. While samples for both Contrast and Result suffer a proportionate decrease in RST-DT, the drop in Narration instances is more substantial. We suggest that there is a second factor at play, namely a genre effect. RST-DT is comprised solely of news texts, while the OntoNotes corpus includes a wider range of genres. Figure 4(a) shows that the primary source of Narration contexts in OntoNotes is the Bible, a significant portion of which is narrative. In fact, around 70% of the Narration contexts are found in the Bible, despite the Bible constituting only 17% of the corpus texts, as shown in Figure 4(b). The next most frequent source of Narration contexts is transcripts of spoken conversations. News articles and other written texts, which make up almost half of the OntoNotes corpus, contribute the least. This pattern confirms the relative infrequency of Narration contexts in news texts, which accounts for their scarcity in the RST-DT corpus. We leave further exploration of this aspect to future work.

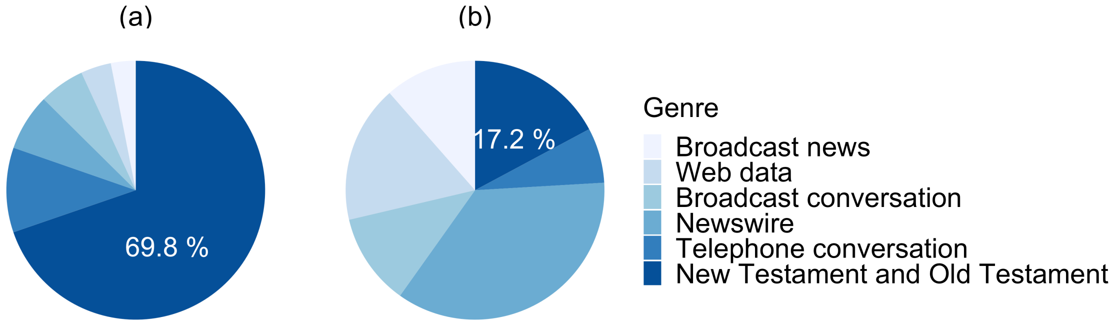


Figure 4: The left figure (a) presents the distribution of Narration coreference samples by genres in OntoNotes. The pie chart on the right (b) shows the genre distribution in OntoNotes.

5 Experimental evaluation of Weak and Strong Bayes: Corpus passage continuation

In our second study, we use passages from the OntoNotes corpus as stimuli in a controlled passage continuation experiment. These selected passages possess two key characteristics: they include connectives as explicit markers of discourse relations and involve animate entities. Using corpus passages makes the sentences more natural than those typically used in psycholinguistic studies, while preserving the control over other variables that may potentially influence pronoun production, such as referent animacy. We will test the findings from the corpus-based analyses, which provide evidence supporting Strong Bayes; and evaluate the predictions of Weak Bayes regarding pronoun interpretation, comparing them with the predictions of other competing models. This requires collecting data on pronoun interpretation, which cannot be obtained solely through corpus analyses.

5.1 Method

Materials and design. We extracted 30 passages from the OntoNotes corpus (Weischedel et al. 2013), where each passage comprised two sentences. The initial/left-hand sentence of each passage depicted an event involving two same-gender human referents in subject and object roles. The subsequent sentence began with an explicit connective, signaling either Narration, Contrast, or Result, as exemplified in Passage (11).

(11) Netanyahu has recently moved ahead of Barak in popularity polls. However, the former Prime Minister can't run in the special vote because he's not currently a member of parliament. –OntoNotes: bn/cnn/01/cnn_0134

To make the referents' gender clear and provide participants with a greater variation in the choice of referring expressions for subsequent re-mentions, we made modifications to the original referring expressions of the two characters in the left-hand sentence of several passages. These modifications fell into three categories: 1) substituting gender-ambiguous expressions with gender-informative ones (e.g., *Netanyahu* → *Benjamin Netanyahu*); 2) replacing pronouns with names or descriptions (e.g., *He* → *Senior U.S official James O'Brien*); and 3) exchanging unusual Biblical names for common English names (e.g., *Nebuchadnezzar* → *Nicolas*).

We then truncated each passage immediately after the connective to create a continuation prompt. Our experimental design incorporated a 3 × 2 factorial structure, employing stimuli analogous to those in (12)–(17). This design varied the Relation Type (Narration, Contrast, Result) and Prompt Type (bare vs. pronoun) within participants and passages.[10] The bare-prompt condition provided measures of next-mention rates and pronoun production rates, which we analyzed for sensitivity to Relation Type and used to compute pronoun interpretation model estimates. The pronoun prompt, on the other hand, provided observed pronoun interpretation biases that we compared with competing pronoun interpretation model estimates.

(12) [Narration, Bare]: Benjamin Netanyahu has recently moved ahead of Ehud Barak in popularity polls. Afterwards, __________

(13) [Narration, Pronoun]: Benjamin Netanyahu has recently moved ahead of Ehud Barak in popularity polls. Afterwards, he __________

(14) [Contrast, Bare]: Benjamin Netanyahu has recently moved ahead of Ehud Barak in popularity polls. However, __________

(15) [Contrast, Pronoun]: Benjamin Netanyahu has recently moved ahead of Ehud Barak in popularity polls. However, he __________

(16) [Result, Bare]: Benjamin Netanyahu has recently moved ahead of Ehud Barak in popularity polls. As a result, __________

(17) [Result, Pronoun]: Benjamin Netanyahu has recently moved ahead of Ehud Barak in popularity polls. As a result, he __________

[10] We manipulated Relation Type by varying the connectives used. Specifically, we employed connectives *and then*, *after that*, *afterwards*, *later*, *next* to signal Narration. For indicating Result, we used connectives *so*, *as a result*, and *therefore*. Additionally, *but* and *however* were used to express Contrast. These connectives were used in previous passage extraction, see Table 5.

In order to conceal the target manipulation, we included 30 filler items extracted from the same corpus. These fillers described events featuring a single human character occupying the subject position, as illustrated in Example (18).

(18) Mr. Nixon was a politician possessing strategic foresight and political courage. —OntoNotes: nw/xinhua/02/chtb_0273

Employing a Latin-Square design, we divided the test stimuli into six distinct lists, ensuring that every item appeared in only one condition per list and all conditions were represented across different items. We pseudo-randomized the test stimuli and fillers by interposing one filler between experimental stimuli, preventing the consecutive occurrence of more than two target items.

Participants. 200 individuals were recruited via the crowd-sourcing platform Prolific,[11] and participated in a 30-minute online experiment.[12] Participation was restricted to native English speakers residing in the United Kingdom. Prior to commencing the experiment, participants gave informed consent form and were prompted to respond to three questions regarding their language background.

Each participant was compensated at a rate of £8 per hour for their involvement in the study. Of the initial pool, 28 participants were excluded due to either non-compliance with instructions or the presence of numerous grammatical errors in their responses. These individuals were replaced by an additional 28 participants, resulting in a final sample of 200 (126 females and 74 males; Mean age = 43, SD = 14.63, Range = 18–74) as originally planned. All participants self-identified as native English speakers, with 21 of them reporting themselves bilingual.[13]

Procedure. The study was presented using the JavaScript library *jsPsych* (De Leeuw 2015; version 7.1.2) and hosted on Pavlovia.[14] Participants were directed to the survey and randomly assigned to one of the six lists. Upon receiving written instructions, they proceeded to engage in a passage continuation task consisting of 60 individual trials, each shown separately and presented one at a time. In each trial, participants were presented with a passage fragment displayed in the center of the browser, followed by a blank text field. They were instructed to type the most natural completion that came to mind in the text field provided immediately after the prompt, with no time constraints for submitting a response. Each participant encountered an equal number of items across the six conditions and was not exposed to any passage more than once.

Upon completing the passage continuation task, participants were directed to annotate their own completions. They were asked to indicate, for each passage, the referent they first mentioned in their response. Four options were provided for this purpose: subject referent, object referent, both referents, and other referent, as illustrated in Figure 5.

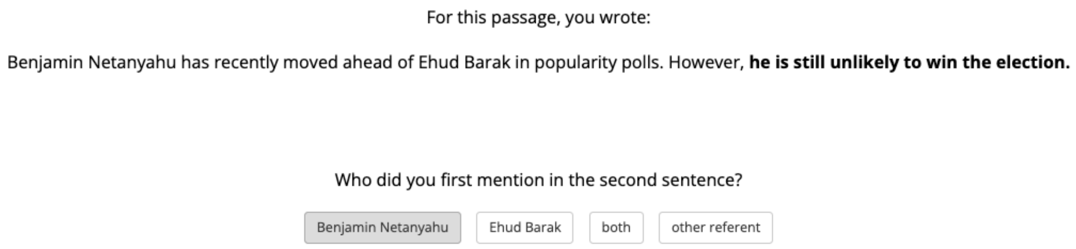
The decision to involve participants in the annotation process was driven by the need to effectively manage the scalability of our study, which included 200 participants and generated 6000 data points in total. As the number of participants increases, so does the workload related to post-experiment manual annotation. By engaging participants in this process, we significantly reduce the manual annotation burden. Additionally, this

[11] URL: https://www.prolific.co/.

[12] The median time spent on the task was 28 minutes.

[13] Rosa & Arnold (2017) observed a predictability effect on pronoun usage, but this was only evident after removing participants who showed little variation in their completions. In line with a reviewer's recommendation, we performed additional analyses, excluding individuals who used fewer than two non-pronouns in their responses. The results are presented in Section C of the Supplementary file.

[14] URL: https://pavlovia.org/.



For this passage, you wrote:

Benjamin Netanyahu has recently moved ahead of Ehud Barak in popularity polls. However, **he is still unlikely to win the election.**

Who did you first mention in the second sentence?

Figure 5: Next-mention annotation in the format of four-choice questions.

approach aids in resolving ambiguities in the annotations more efficiently. Participants, having provided the responses themselves, are better positioned to provide accurate annotations. We believe this method is especially beneficial in large-scale studies, where both the volume of data and the potential for ambiguous interpretations are higher.

Coding. For the analysis, every response was coded for (i) the referent that participants first mentioned in their completion (the previous subject referent, object referent, both subject and object, or other referents), (ii) the referential form of their first mention (pronoun, non-pronoun, or zero). For the former, we examined participants' self-annotations collected during the online experiment. For the latter, we used *spaCy* (Honnibal & Montani 2017), a library for Natural Language Processing in Python, to automatically detect the subject of the participants' completions and label the choice of referring expression based on part-of-speech tags.

Responses were excluded if (a) participants referred to more than one character (e.g., *Peter criticized Jeffery for saying these things. After that, **they** made up.*); (b) participants referred to other entities that were not the main characters in the context sentence (e.g., *Benjamin Netanyahu has recently moved ahead of Ehud Barak in popularity polls. However, **the polls** have proven unreliable in recent years.*); (c) the first mention was elliptical or zero (e.g., *Adam refused to stop chasing Leo. Shortly afterwards, **Ø** became tired and could run no further.*).[15] Approximately 23.5% of trials (out of 6000 trials) were excluded for one of these reasons, resulting in a final dataset of 4588 responses for analysis. Among these responses, 1773 belong to the bare-prompt condition, while 2815 belong to the pronoun-prompt condition.[16]

We implemented a quality control process to ensure the reliability of participants' self-annotations on the next-mention choice. A random sample comprising 10% of the data (458 responses) was subjected to independent coding by two annotators. The two annotators agreed in 96.7% of observations, and the inter-annotator agreement rate for referent choice, as measured by Cohen's kappa (unweighted), was 0.76 ($z = 5.7$, $p < 0.001$). In cases of disagreement, a third annotator was consulted, and through subsequent discussions, a consensus was reached. Following this procedure, we estimated the participants' annotation accuracy to be 93.4%.

[15] To ensure that the exclusion of responses with a zero subject did not bias our results, we examined the proportion of zero subjects across each of the three relations. The rate of zeros did not vary by coherence relation ($\chi^2(2) = 5.52$, $p = 0.06$). The rate of null subjects was 8.3%, 10.5%, and 8.6% in Narration, Contrast, and Result, respectively.

[16] More data were excluded from the bare-prompt condition compared to the pronoun-prompt condition due to the larger variety of referents chosen when there is no pronoun, leading to more references to non-main characters.

To assess the quality and reliability of the automatic labeling process for referential form within bare-prompt completions, again a random sample of responses was selected for quality control. This time, the selection comprised 10% (177 responses) of the relevant bare-prompt data. A single annotator manually examined these responses, as the task was straightforward and objective, with minimal ambiguity involved. We estimated the accuracy of the automatic labeling of referential form to be 98.9%. We considered these levels of accuracy to be acceptable, as they were unlikely to introduce any significant bias into our results due to a higher error rate in one condition compared to another.

5.2 Analysis

The data were analyzed utilizing mixed-effect logistic regressions with centered predictors. We built three models: (1) one focused solely on the bare-prompt data, with the dependent variable being whether the continuation continued with the subject referent or not and a fixed effect for the 3-level discourse relation type; (2) one where the dependent variable assessed whether the next mention was pronominalized, incorporating fixed effects for discourse relation types, grammatical role types (subject or non-subject), and their interaction; and (3) another that considered both the bare-prompt and pronoun-prompt data, using the same dependent variable as the first model but including fixed effects for discourse relation types, prompt types (bare or pronoun), and their interaction. Model building began with a maximal random structure and subsequently simplifying the random-effects structure until convergence was attained (Barr et al. 2013). To assess the significance of the fixed effects, we employed likelihood ratio tests, comparing the full model with the effect in question against a counterpart model devoid of said effect.

The fitting of these models was conducted using the *lme4* package (Bates et al. 2015) in *R* (version 4.1.2, R Core Team 2021). For each of the variables in the model, we report the coefficients in log odds. Null-hypothesis significance testing was employed to ascertain the statistical significance of the results (alpha level: 0.05).

In instances where the *lme4* package encountered convergence failure, we employed a Bayesian approach using the *brms* R package (Bürkner 2017). To ensure stable inferences, we utilized weakly informative priors, specifically Cauchy distributions with a center of 0 and a scale of 2.5 (Gelman et al. 2013); and a maximal random structure. All fits were run with six chains, each comprising 2000 iterations, with half as warm-up. Prior to analysis, thorough diagnostic checks were conducted to rule out any potential pathologies in the estimation process. For the Bayesian models, we reported the estimated mean and the corresponding 95% Credible Intervals (CIs) of the posterior distribution in log odds. The 95% CI represents the range within which the outcome is likely to fall with a 95% probability, based on the observed data. A null hypothesis is rejected if the interval does not include zero (Gelman et al. 2013).

5.3 Results

In Section 3, we outlined our research questions and hypotheses. We state them again here. Our first question asks which model for pronouns (i.e., Bayesian, Expectancy, or Mirror) best accounts for the observed interpretation biases. Additionally, we have put forth two hypotheses regarding next-mention biases. Hypothesis 1 posits that, in Narration contexts, the likelihood of the subject referent being mentioned again is higher than in Contrast or Result contexts. If our data supports Hypothesis 1, it leads us to Hypothesis 2, which concerns pronoun production biases. Specifically, we hypothesize that discourse relations do not influence pronoun production biases, aligning with the predictions of the Strong Bayes model.

Next-mention biases. Raw proportions of subject references by discourse relation and prompt type are shown in Figure 6. We first evaluated the next-mention biases (red bars) in the bare-prompt condition. Analyses of the binary outcome of subject coreference showed a main effect of Relation Type ($\chi^2(2) = 22.48$, $p < .001$). Further pairwise comparisons revealed that Narration relations yield the most subject continuations, more than Contrast (z $= 4.29$, $p < .001$) and Result (z $= 3.73$, $p < .001$) in the bare-prompt condition; no difference was found between Contrast and Result (z $= 0.06$, $p \approx 1$). A model summary is presented in Table 11. Therefore, Hypothesis 1 receives support from the data: subject referents are more predictable in Narration than in Contrast and Result.

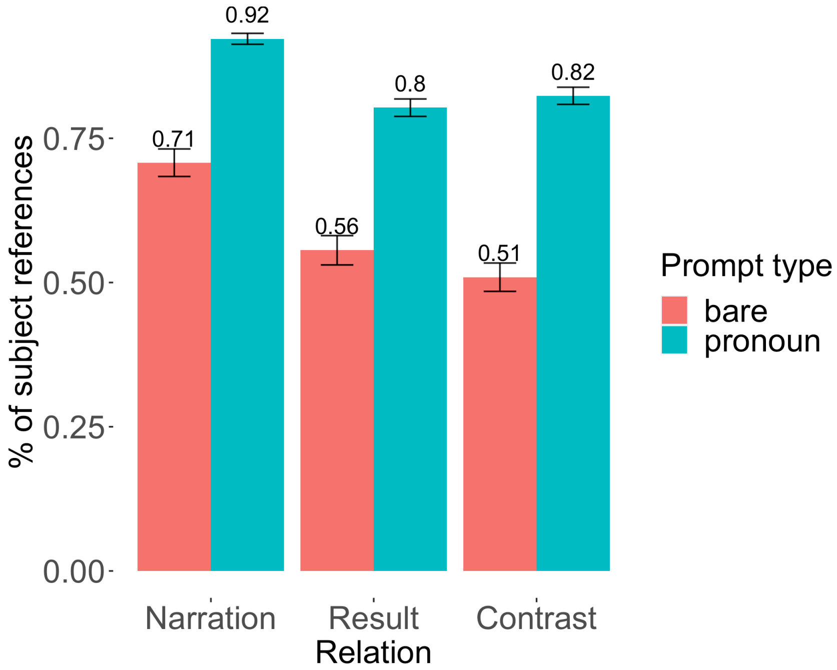


Figure 6: Proportion of subject references by discourse relations and prompt types. Subject references in the bare-prompt condition represent participants' expectations about the subject being the next mention, while subject references in the pronoun-prompt condition allow us to measure participants' interpretation biases when encountering an ambiguous pronoun. Error bars are standard errors over by-participant means.

Fixed effects	Estimate	SE	Z	*p*
Intercept	0.64	0.25	2.58	
Narration	**0.63**	**0.13**	**4.68**	**<0.001**
Result	**−0.32**	**0.15**	**−2.12**	**0.03**

Table 11: Summary of logit mixed effect models of next mention with a fixed effect for the 3-level discourse relation type.

Pronoun production biases. The bare-prompt condition also allows us to measure participants' pronominalization rates (see Figure 7). To model the binary outcome of whether the participant produced a pronoun or not, we built a full Relation Type × Grammatical Role model and replicated the well-known effect of grammatical role on pronoun production, with more pronouns produced referring to subjects than non-subjects (β = 2.71, [2.11, 3.40]). As for Hypothesis 2, we found no evidence of an effect of Relation Type on pronominalization, in keeping with the findings from our corpus-based observational analyses. A model summary is presented in Table 12.

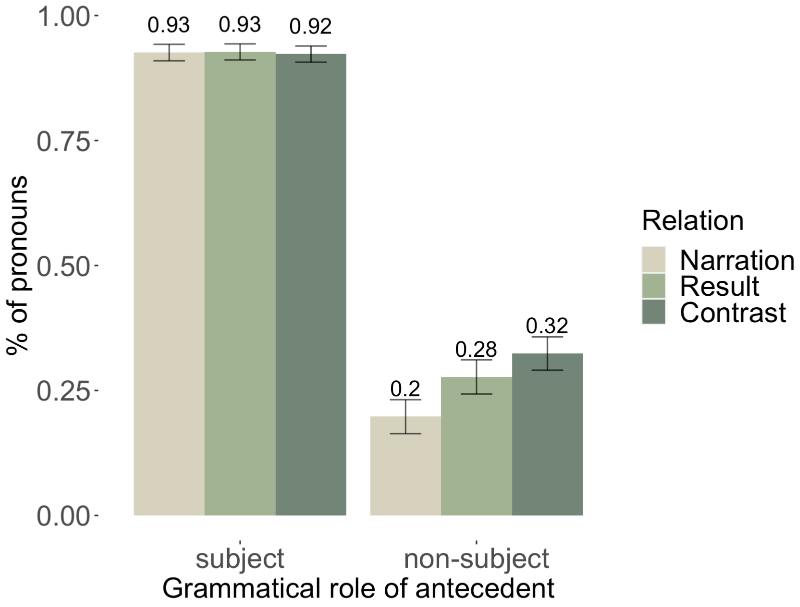


Figure 7: Pronominalization rates by grammatical roles and relation types.

	Estimate	Est.Error	95% CI
Intercept	1.05	0.30	[0.47, 1.67]
Narration	−0.04	0.24	[−0.50, 0.44]
Result	0.18	0.21	[−0.22, 0.61]
subject	**2.71**	**0.32**	**[2.11, 3.40]**
Narration:subject	0.36	0.27	[−0.15, 0.93]
Result:subject	−0.32	0.24	[−0.79, 0.15]

Table 12: Summary of logit mixed effect models of pronoun production (with all predictors centered).

Beyond traditional hypothesis testing, we further evaluate evidence in favor of the null hypothesis (i.e., the absence of the effect) using Bayes Factors, similar to our approach in the corpus analyses discussed in Section 4.3.

We compared two Bayesian models: Model 1 (H1: alternative hypothesis) and Model 2 (H0: null hypothesis). Model 1 incorporates both *Relation Type* and *Grammatical Role* as predictors, as well as their interaction, same as the fixed effect items in the model presented in Table 12.[17] Model 2, in contrast, excluded *Relation Type* and considered

[17] We simplified the maximal random effect structure in both Model 1 and Model 2 to include only the random intercepts for participants and items in order to facilitate the estimation process of Bayes Factors.

only *Grammatical Role* as a predictor. The Bayes Factor in favor of Model 2 (H0) over Model 1 (H1) was estimated to be 36432.67 (in favor of Model 1 over Model 2: 0.00003), indicating strong evidence in support of Model 2. Therefore, our analysis indicates that the inclusion of *Relation Type* in the model does not contribute significantly to explaining pronoun production biases, and the simpler Model 2, representing the null hypothesis or Strong Bayes, is the preferable model. This is consistent with the finding in our corpus analyses.

Pronoun interpretation biases. Next, we compare participants' interpretation biases measured in the pronoun-prompt condition (depicted in the blue bars of Figure 6) to their next-mention biases (already seen in the red bars of Figure 6). To do this, we constructed a mixed-logit model of the binary outcome of subject versus non-subject continuation, incorporating the fully crossed factors of Relation Type × Prompt Type. The summary of this model is presented in Table 13.

We find a main effect of Relation Type ($\chi^2(2) = 45.4$, $p < .001$) whereby Narration yields the most subject continuations, more than Contrast (z = 6.52, $p < .001$) and Result (z = 5.79, $p < .001$); no difference was found between Contrast and Result (z = 0.08, $p \approx 1$). We also replicate the well-known effect of Prompt Type ($\chi^2(1) = 72.5$, $p < .001$) whereby the presence of a pronoun increases subject continuations; the interpretation bias is more skewed towards the subject in this condition than in the bare-prompt condition. The mean score for interpreting an ambiguous pronoun as subject in Narration was significantly higher than the score in Contrast ($z = 6.37$, $p < .001$), and in Result ($z = 5.85$, $p < .001$). No difference was found between Contrast and Result ($z = 0.01$, $p \approx 1$).

The interaction between Prompt Type and Relation Type shows that Prompt Type, changing from the bare-prompt condition to the pronoun-prompt condition, has the biggest effect in increasing subject continuations in the Narration condition. Although this initially appears counter-intuitive when comparing the raw proportions of differences between the bare and pronoun prompts (21% in Narration relations versus 31% in Contrast and 24% in Result relations), it can be better understood when considering the following perspective. In the bare-prompt condition, subject references in Narration relations sit high at 71%, limiting further increase. Transitioning to the pronoun-prompt escalates this to 92%, nearing saturation. This marginal potential for growth heightens the interaction effect.

Fixed effects	Estimate	SE	Z	*p*
Intercept	1.59	0.23	6.87	
Narration	**0.81**	**0.11**	**7.09**	**<0.001**
Result	**−0.41**	**0.12**	**−3.67**	**<0.001**
pronoun prompt	**0.97**	**0.11**	**8.82**	**<0.001**
Narration:pronoun prompt	**0.21**	**0.08**	**2.76**	**0.006**
Result:pronoun prompt	−0.10	0.07	−1.44	0.15

Table 13: Summary of logit mixed effect models of next mention with the fully crossed factors of Relation Type × Prompt Type (with all predictors centered).

5.4 Quantitative model comparisons

To address our research question on model evaluation, as formulated in Section 3, we conduct a quantitative analysis to assess the performance of three models by comparing their predictions against the observed interpretation biases: Bayes, Expectancy (relying primarily on the next-mention bias), and Mirror Model (based on a claim that speakers use pronouns when referring to entities whose salience makes them the preferred referent for a listener). The models are formalized in Table 14.

(a) Bayes: $$P(\text{referent} \mid \text{pronoun}) = \frac{P(\text{pronoun}|\text{referent})\,P(\text{referent})}{\sum_{\text{referent}\in\text{referents}} P(\text{pronoun}|\text{referent})\,P(\text{referent})}$$

(b) Mirror: $$P(\text{referent} \mid \text{pronoun}) \leftarrow \frac{P(\text{pronoun}|\text{referent})}{\sum_{\text{referent}\in\text{referents}} P(\text{pronoun}|\text{referent})}$$

(c) Expectancy: $$P(\text{referent} \mid \text{pronoun}) \leftarrow P(\text{referent})$$

Table 14: Formulations of the three models of pronoun interpretation.

It follows from the formalization that, to determine model predictions, we need to estimate two probabilities for each of the 90 experimental stimuli (30 items × 3 relations): next-mention biases, $P(\text{referent})$, and pronoun production biases, $P(\text{pronoun}|\text{referent})$. These quantities are estimated based on the bare-prompt condition of our experiment, wherein participants have the freedom to select both the referent and the form employed to refer to the referent. To prevent zero-probability estimates, which could result in undefined model predictions for certain stimuli, we use simple additive smoothing (Schutze et al. 2008). We add a pseudo-count of one to our stimulus-specific experimental data for each logically possible combination of the V = 2 referents (subject and non-subject referents) and the W = 2 forms (pronoun, non-pronoun) that could be used in a re-mention. This approach yields stimulus-specific probability estimates as follows:

(IV) $$\hat{P}(NP_j) = \frac{Count(NP_j) + W}{Count(NP1) + Count(NP2) + V \times W}$$

(V) $$\hat{P}(\text{pronoun}|NP_j) = \frac{Count(NP_j \wedge \text{pronoun}) + 1}{Count(NP_j) + W}$$

Following Zhan et al. (2020), we computed stimulus-by-stimulus predictions of the three models for pronoun interpretation preferences, and compared them against stimulus-by-stimulus observed behavior in the pronoun-prompt condition. As in Zhan et al. (2020), we used three statistical metrics to conduct a comprehensive evaluation of how well the predictions of these models align with the observed interpretation biases. These metrics, each focusing on a different aspect of model performance, include: R-squared (R^2), Mean Squared Error (MSE), and Average Cross Entropy (ACE). We explain them below.

R^2 measures the proportion of the variance in the observed interpretation biases that can be explained by the predicted interpretation preferences of each model in a linear regression model. It is used to gauge the goodness of fit of these models. The R^2 value ranges from 0 to 1, where a value closer to 1 indicates a stronger fit. However, this metric may still indicate a high fit even when the model predictions are poorly calibrated (if the model systematically underestimates or overestimates the observed biases), hence additional metrics are necessary to assess the models' performance.

We use Mean Squared Error (MSE), which calculates the average squared difference between the predicted and observed values. It is employed to estimate the average magnitude of the errors made by the models. Lower MSE values are indicative of a better model fit and reduced prediction errors. It is defined as (VI), where y_i and $\hat{y}_i$ are respectively the observed and predicted interpretation preferences for the i-th stimulus and n is the total number of stimuli.

(VI)
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Similar to MSE, the Average Cross Entropy (ACE), defined as (VII), quantifies the dissimilarity between the observed rate y_i and the predicted rate $\hat{y}_i$. While MSE weights the discrepancies equally throughout all items, ACE more heavily penalizes predictions that are close to 0 when the observed rate is actually close to 1, as well as predictions that are near 1 when the observed rate is close to 0. In other words, ACE assigns greater importance to evaluating extreme cases.

(VII)
$$\text{ACE} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \cdot \log_2(\hat{y}_i) + (1 - y_i) \cdot \log_2(1 - \hat{y}_i)\right)$$

Hence, when evaluating three models for pronoun interpretation, a better model is indicated by a higher R-squared (R^2) value and lower Mean Squared Error (MSE) and Average Cross Entropy (ACE) scores.

Results. The results from all three metrics suggest that the Bayes Model outperforms the two competing models, as displayed in Table 15.[18] Mirror Model demonstrates inferior performance when evaluated using the R^2 metric. However, when the evaluation is based on the MSE and the ACE metrics, Expectancy Model exhibits comparatively weaker performance.

	Bayes	Expectancy	Mirror
R^2	**0.50**	0.43	0.03
MSE	**0.03**	0.10	0.06
ACE	**0.58**	0.82	0.71

Table 15: Results of statistical metrics for model comparisons. Best results boldfaced.

Figure 8 illustrates the observed pronoun interpretation rates against stimulus-specific predictions from each model separately in three distinct plots, with the dotted x = y line representing an ideal model fit. A significant number of predictions cluster above the perfect-fit dotted line, suggesting that the models tend to underestimate subject interpretation. The Bayes Model shows a noteworthy performance particularly when the human biases towards interpreting pronouns as referring to the subject are approaching 1.

In addition to the metrics from Zhan et al. (2020), we expanded our evaluation of the three pronoun interpretation models by incorporating the Bayesian method from Patterson et al. (2022). The results obtained using this Bayesian method were consistent with those using the metrics of Zhan et al. (2020). For more details on the use of the Bayesian method, see Section D of the Supplementary file.

[18] We evaluated the models based solely on their predictions for subject referents. Including non-subject referents did not affect the MSE score, but it resulted in higher R-squared (R^2) scores for all three models: 0.81 (Bayes), 0.62 (Mirror), 0.38 (Expectancy). This is because the additional variation between subject and non-subject referents in the observed pronoun interpretation biases could be accounted for by the models.

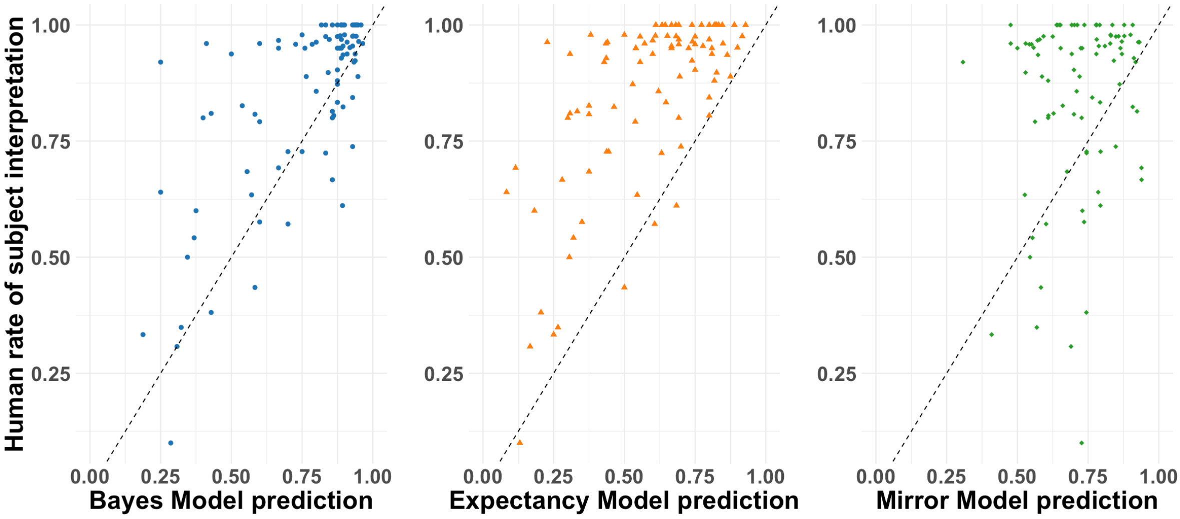


Figure 8: Quantitative model evaluation for subject referents only. The figure consists of three separate plots, each showing the predicted pronoun interpretation rates from a different model. In total, there are 270 datapoints. Each datapoint represents a prediction made by each model for each of the 90 stimuli (30 items × 3 relations in the pronoun-prompt condition).

6 Discussion

6.1 Summary

We used both corpus-based and experimental methodologies to assess whether the Bayesian probabilistic framework adequately captures pronoun production and interpretation, and the relationship between them in naturalistic contexts. According to this framework, listeners utilize Bayesian principles to reverse-engineer the intended referent of a speaker. Additionally, a strong form of this framework proposes that next-mention bias and pronoun production bias are influenced by distinct contextual factors.

Our observational corpus analyses reveal that while the frequency of subject re-mentions varies across distinct discourse relations (Narration, Contrast, and Result), the rate of pronominalization remains consistent, in line with the prediction of the Strong Bayes. This observation is consistent irrespective of whether relations are explicitly signaled by connectives or inferred by human coders.

To further investigate these findings, we conducted a passage completion experiment utilizing corpus passages as stimuli. Manipulating discourse relation (Narration, Contrast, and Result) and prompt type (bare and pronoun), we obtained empirical evidence that favors the Bayesian Model over two competing models in predicting the observed pronoun interpretation bias. Moreover, our results once again highlight a separation between the factors influencing next-mention bias and those influencing pronoun production bias, in line with the prediction of the Strong Bayes.

In summary, our data provide broad support for the predictions of the Bayesian Model, both in its weak and strong forms. We discuss them in more detail below.

6.2 Weak Bayes

The central claim of the Bayesian Model, in its weaker form, is that the relationship between pronoun production and interpretation biases can be modelled through Bayes' theorem. This proposition asserts that listeners, upon encountering a pronoun, undertake the task of reverse-engineering the speaker's targeted referent using Bayesian mechanisms, as formulated in Eq. II. Interpretation biases thus should reflect both next-mention biases and production biases. This was substantiated by our experimental findings, where the variation pattern in interpretation biases is the same as that in next-mention biases, suggesting that the variation in next-mention biases shaped by the discourse relation manipulation became apparent through their influence on interpretation biases. Moreover, interpretation biases are much more skewed towards the previous subject across relations, in contrast to the corresponding next-mention biases, indicating the expected effect of pronoun production biases.

Further, we analyzed the predictions of the Bayesian Model along with two other models - the Expectancy Model and the Mirror Model. The bare-prompt condition data was employed to estimate the predictions of these three models, and three metrics —R^2, MSE, and ACE —were used to evaluate model performance. The Bayesian Model typically outperformed both rival models. In general, the Expectancy Model underestimated the bias towards interpreting a pronoun as referring to the previous subject. This tendency is visible in Figure 8, where Expectancy Model points are above the x = y perfect-prediction line: this is due to the Expectancy Model not including a term for pronoun production, which biases pronouns towards previous-subject interpretations. In contrast, the Mirror Model often underestimated cross-stimuli variability in interpretation preferences, observable in Mirror Model points clustering surrounding the x = 0.75, regardless of the actual stimulus-specific interpretation bias. This pattern is due to the Mirror Model disregarding the effect of next-mention bias, which shows more variability across stimuli than the pronoun production bias.

On the other hand, the alignment between Bayesian Model predictions and observed interpretation rates isn't perfect either. The Bayesian Model tended to underestimate the interpretation bias towards the previous subject, though not as much as the Expectancy Model. We note that in 16 out of 90 stimuli (10 Occasion, 3 Result, and 3 Contrast), the observed interpretation rate reaches the maximum limit of 1, implying that all human participants interpreted the pronoun as referring to the previous subject. However, after applying our smoothing method, the predicted interpretation rate by all three models is unlikely to reach 1, rendering a perfect match unattainable in these instances. Bayesian Model predictions tend to be concentrated in the upper right corner, suggesting superior performance in these extreme cases, especially in Occasion scenarios. In fact, in 26 out of 30 Occasion scenarios, the observed interpretation bias towards the previous subject exceeded 0.90, while the proportions were 14 out of 30 for Contrast, and 12 out of 30 for Result. Therefore, with our stimuli, there was a clear inclination for participants to interpret the ambiguous pronoun as the subject. This can be attributed to the fact that our stimuli of discourse relations typically represent natural discourse where subject continuity is much more common. Scatter points in Figure 8 would less likely be gathered in the upper right corner if we used object-biased implicit causality verbs (such as *admire*) or transfer-of-possession verbs (like *give*) in our stimuli.

In the context of these observations, it's worth noting that previous studies (Rohde & Kehler 2014; Zhan et al. 2020; Patterson et al. 2022) have consistently found that the Bayesian Model of pronoun interpretation outperforms alternative models across diverse languages and experimental setups. Our study, while employing a different approach by using corpus passages as stimuli instead of constructed material, aligns with this prevailing trend. By bridging the gap between constructed material and more naturalistic language, our findings contribute to the growing body of evidence that supports the role of Bayesian inference in pronoun interpretation.

6.3 Strong Bayes

Our results from both observational and experimental studies are in keeping with the prediction of Strong Bayes that the likelihood of next mention and the likelihood of pronoun production are conditioned by different sets of factors. Specifically, we show that the likelihood of re-mentioning the subject is influenced by factors related to discourse coherence, while the likelihood of pronoun production appears to be insensitive to differences between discourse relations, but primarily subject to grammatical role, whereby subject re-mentions were pronominalized significantly more often than non-subject rementions.

In the literature, this question is frequently reframed as an investigation into whether the production of pronouns is influenced by the referent predictability induced by a set of semantico - pragmatic factors. Assuming that addressees can (to some extent) predict which referent will be mentioned next, speakers could exploit addressees' expectations: They could use less costly expressions, like pronouns, more frequently when referents are more expected or predictable to their addressees. This exploitation of predictability is often associated with the view of language as an efficient code for communication (Tily & Piantadosi 2009). In fact, predictability has been broadly shown to account for reduction processes in many areas of language production, such as attenuated pronunciations for more predictable words (e.g. Jurafsky et al. 2001).

Contradicting this perspective, our findings in this task show that speakers do not use more pronouns in Narration scenarios where the subject referent is more predictable. The fact that speakers do not exploit predictability is *prima facie* counter-intuitive because it looks like it makes them less efficient. One explanation may be that, when speakers produce language, their cognitive load is such that they can only rely on comparatively shallower cues (such as grammatical role) rather than combining them with predictability. Therefore, speakers might ignore cues that their addresses are sensitive to. Another potential reason is that the influence of predictability is so minor that it becomes eclipsed by other factors, such as the grammatical role or topichood, which is the main driver of a speaker's choice of referential form.

6.4 Methodological implications

From a methodological perspective, our exploratory analyses showcase both the advantages and the challenges of using a corpus-based approach to assess the prediction of Strong Bayes.

The clearest advantage of corpus-based approaches is the fact that they allow the use of naturally produced language, and the examination of varied and diverse contexts. Most of the previous studies instead focused on a few verb types and used carefully controlled contexts to elicit continuations. Another advantage of our approach is that, unlike most of the previous corpus work (e.g. Arnold 2001; Guan & Arnold 2021), we rely on co–reference annotated corpora to enable automatic data extraction, which saves tremendous human efforts compared to manual annotation.

At the same time, using corpus data also poses some challenges. When the objective is to re-examine experimental results, as in our analyses with verb types, the difference between controlled contexts and naturally occurring language can get in the way. Future research on study replication could look for genres that may have more similar uses. For instance, for transfer-of-possession scenarios, one could use a corpus of football commentary that is abundant in expressions describing dynamic ball passing.

Another challenge is the fact that the stylistic features of a given corpus might not be tailored to the contexts of interest. For instance, in our analysis with manually annotated discourse relations, we found contexts of Narration to be fairly limited because the RST-DT corpus is fully composed of news articles, where narrative relations are relatively scarce.

To conclude, our study provides evidence for the generality of the Bayesian Model and contributes to the field in two ways. Methodologically, we have exemplified how linguistic research can benefit from resources developed in Computational Linguistics, in particular co-reference annotated corpora. We hope that our work will spark interest in the use of such resources to address open questions in theoretical linguistics. Our study is also the first to extend the empirical base to more naturalistic corpus passage completions.

At a theoretical level, we show that the Bayesian Model overall made more accurate predictions for pronoun interpretation than production or next-mention biases separately; we also show that discourse relations between clauses exhibit systematic patterns regarding next-mention bias, which again broadens the empirical scope of the debate; we find evidence consistent with the prediction of Strong Bayes that next-mention bias and pronoun production bias are influenced by distinct sets of factors.

Data availability

All data processing and analysis code are publicly accessible through an OSF repository: https://doi.org/10.17605/OSF.IO/N54SW. The supplementary file is available in the same OSF repository, which includes

i. Section A: More information in the corpus analyses with verb types
ii. Section B: More information in the analyses with discourse relations
iii. Section C: Analysis results excluding participants with low variation in referring expressions
iv. Section D: Comparison of pronoun interpretation models using Bayesian methods

Ethics and consent

This study was approved by the PPLS Research Ethics Committee of the University of Edinburgh (approval no. 322-2122/4). Informed consent was obtained from all individual participants included in the study.

Funding information

This research is part of the project EXPEDIS (PID2021-122779NB-I00) funded by MICIU/AEI/10.13039/501100011033 and by “ERDF A way of making Europe”.

Acknowledgements

We thank Professor Bruno Nicenboim for his help with the Bayesian method used to compare the three models. Our thanks also go to the audiences from Sinn und Bedeutung 26 and Human Sentence Processing 23, where earlier versions of this research were presented, as well as to our colleagues at the Department of Translation and Language Sciences at Universitat Pompeu Fabra and the School of Philosophy, Psychology & Language Sciences at the University of Edinburgh for their questions and comments. Finally, we wish to express our sincere appreciation to the anonymous reviewers and Associate Editor Sol Lago for their constructive comments and suggestions.

Competing interests

The authors have no competing interests to declare.

References

Ariel, Mira. 1990. *Accessing noun phrase antecedents*. Londres: Routledge.

Ariel, Mira. 2001. Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects* 8. 29–87. https://doi.org/10.1075/hcp.8.04ari

Arnold, Jennifer E. 1998. *Reference form and discourse patterns*. Stanford: Stanford University dissertation.

Arnold, Jennifer E. 2001. The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse processes* 31(2). 137–162. https://doi.org/10.1207/S15326950DP3102_02

Arnold, Jennifer E. 2010. How speakers refer: The role of accessibility. *Language and Linguistics Compass* 4(4). 187–203. https://doi.org/10.1111/j.1749-818X.2010.00193.x

Arnold, Jennifer E. & Brown-Schmidt, Sarah & Trueswell, John. 2007. Children's use of gender and order-of-mention during pronoun comprehension. *Language and cognitive processes* 22(4). 527–565. https://doi.org/10.1080/01690960600845950

Arnold, Jennifer E. & Tanenhaus, Michael K. 2011. Disfluency effects in comprehension: How new information can become accessible. In Gibson, Edward A. & Pearlmutter, Neal J. (eds.), *The Processing and Acquisition of Reference*, 197–217. The MIT Press. https://doi.org/10.7551/mitpress/9780262015127.003.0008

Asher, Nicholas & Lascarides, Alex. 2003. *Logics of conversation*. Cambridge: Cambridge University Press.

Asher, Nicholas & Vieu, Laure. 2005. Subordinating and coordinating discourse relations. *Lingua* 115(4). 591–610. https://doi.org/10.1016/j.lingua.2003.09.017

Bader, Markus & Portele, Yvonne. 2019. The interpretation of German personal pronouns and d-pronouns. *Zeitschrift für Sprachwissenschaft* 38(2). 155–190. https://doi.org/10.1515/zfs-2019-2002

Barr, Dale J & Levy, Roger & Scheepers, Christoph & Tily, Harry J. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68(3). 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, Douglas & Mächler, Martin & Bolker, Ben & Walker, Steve. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. https://doi.org/10.18637/jss.v067.i01

Brennan, Susan E. 1995. Centering attention in discourse. *Language and Cognitive processes* 10(2). 137–167.

Bürkner, Paul-Christian. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software* 80. 1–28. https://doi.org/10.18637/jss.v080.i01

Carlson, Lynn & Marcu, Daniel. 2001. Discourse tagging reference manual. Tech. Rep. ISI-TR-545 University of Southern California Information Sciences Institute.

Carlson, Lynn & Okurowski, Mary Ellen & Marcu, Daniel. 2002. *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania. https://doi.org/10.35111/4w31-m996

Crawley, Rosalind A & Stevenson, Rosemary J & Kleinman, David. 1990. The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research* 19(4). 245–264.

De Leeuw, Joshua R. 2015. jspsych: A Javascript library for creating behavioral experiments in a Web browser. *Behavior research methods* 47. 1–12. https://doi.org/10.3758/s13428-014-0458-y

Ferretti, Todd R & Rohde, Hannah & Kehler, Andrew & Crutchley, Melanie. 2009. Verb aspect, event structure, and coreferential processing. *Journal of memory and language* 61(2). 191–205. https://doi.org/10.1016/j.jml.2009.04.001

Ferstl, Evelyn C & Garnham, Alan & Manouilidou, Christina. 2011. Implicit causality bias in English: A corpus of 300 verbs. *Behavior Research Methods* 43(1). 124–135. https://doi.org/10.3758/s13428-010-0023-2

Frank, Stefan L. & Otten, Leun J. & Galli, Giulia & Vigliocco, Gabriella. 2013. Word surprisal predicts N400 amplitude during reading. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 878–883. Sofia, Bulgaria: Association for Computational Linguistics. https://aclanthology.org/P13-2152.

Frederiksen, Anne Therese & Mayberry, Rachel I. 2022. Pronoun production and comprehension in American Sign Language: the interaction of space, grammar, and semantics. *Language, Cognition and Neuroscience* 37(1). 80–102. https://doi.org/10.1080/23273798.2021.1968013

Fukumura, Kumiko & Van Gompel, Roger. 2010. Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language* 62(1). 52–66. https://doi.org/10.1016/j.jml.2009.09.001

Gelman, Andrew & Carlin, John & Stern, Hal & Dunson, David & Vehtari, Aki & Rubin, Donald. 2013. *Bayesian data analysis*. New York: Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Givón, Talmy. 1983. *Topic continuity in discourse*. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.3

Gronau, Quentin F. & Singmann, Henrik & Wagenmakers, Eric-Jan. 2020. bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software* 92(10). 1–29. https://doi.org/10.18637/jss.v092.i10

Guan, Shuang & Arnold, Jennifer E. 2021. The predictability of implicit causes: testing frequency and topicality explanations. *Discourse Processes* 1–27. https://doi.org/10.1080/0163853X.2021.1974690

Gundel, Jeanette K. 1988. Universals of topic-comment structure. *Studies in syntactic typology* 17(1). 209–239. https://doi.org/10.1075/tsl.17.16gun

Gundel, Jeanette K & Hedberg, Nancy & Zacharski, Ron. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 274–307. https://doi.org/10.2307/416535

Hobbs, Jerry R. 1990. *Literature and cognition*. Stanford: Center for the Study of Language (CSLI).

Hoek, Jet & Kehler, Andrew & Rohde, Hannah. 2021. Pronominalization and expectations for re-mention: Modeling coreference in contexts with three referents. *Frontiers in Communication* 6. 674126.

Holler, Anke & Suckow, Katja. 2016. How clausal linking affects noun phrase salience in pronoun resolution. In Suckow, Anke Holler Katja (ed.), *Empirical perspectives on anaphora resolution*, 61–85. Berlin: de Gruyter. https://doi.org/10.1515/9783110464108-005

Honnibal, Matthew & Montani, Ines. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Hwang, Heeju. 2023a. Choice of nominative and topic markers in Korean discourse. *Quarterly Journal of Experimental Psychology* 76(4). 905–921. https://doi.org/10.1177/17470218221103544

Hwang, Heeju. 2023b. The influence of discourse continuity on referential form choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 49(4). https://doi.org/10.1037/xlm0001166

Hwang, Heeju & Lam, Suet Ying. 2023. The influence of action continuity on reference form in Mandarin and English. Poster presented at 36th Annual Conference on Human Sentence Processing, University of Pittsburgh.

Hwang, Heeju & Lam, Suet Ying & Ni, Wenjing & Ren, He. 2022. The role of grammatical role and thematic role predictability in reference form production in Mandarin Chinese. *Frontiers in Psychology* 13. https://doi.org/10.3389/fpsyg.2022.930572

Jurafsky, Daniel & Bell, Alan & Gregory, Michelle & Raymond, William D. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. *Typological studies in language* 45. 229–254. https://doi.org/10.1075/tsl.45.13jur

Kass, Robert & Raftery, Adrian. 1995. Bayes factors. *Journal of the American Statistical Association* 90(430). 773–795. http://www.jstor.org/stable/2291091.

Kehler, Andrew. 2002. *Coherence, reference, and the theory of grammar*. Stanford: CSLI publications.

Kehler, Andrew. 2019. Coherence Relations. In Truswell, Robert (ed.), *The Oxford Handbook of Event Structure*, Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199685318.013.27

Kehler, Andrew & Kertz, Laura & Rohde, Hannah & Elman, Jeffrey L. 2008. Coherence and coreference revisited. *Journal of Semantics* 25(1). 1–44. https://doi.org/10.1093/jos/ffm018

Kehler, Andrew & Rohde, Hannah. 2013. A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics* 39(1-2). 1–37. https://doi.org/10.1515/tl-2013-0001

Kehler, Andrew & Rohde, Hannah. 2019. Prominence and coherence in a Bayesian theory of pronoun interpretation. *Journal of Pragmatics* 154. 63–78. https://doi.org/10.1016/j.pragma.2018.04.006

Konuk, Gökben & von Heusinger, Klaus. 2021. Discourse prominence in Turkish: The interaction of grammatical function and semantic role. In Khomchenkova, Irina & Sinitsyna, Yulia & Tatevosov, Sergei (eds.), *Proceedings of the 15th workshop on Altaic formal linguistics*. 109–120.

Kravtchenko, Ekaterina. 2022. *Integrating pragmatic reasoning in an efficiency-based theory of utterance choice*. Saarbükren: Universität des Saarlandes dissertation.

Lam, Suet-Ying & Hwang, Heeju. 2022. How does topicality affect the choice of referential form? Evidence from Mandarin. *Cognitive Science* 46(10). e13190. https://doi.org/10.1111/cogs.13190

Lambrecht, Knud. 1996. *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*, vol. 71. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511620607

Lindemann, Sofiana-Iulia & Mada, Stanca & Sasu, Laura & Matei, Madalina. 2020. Thematic role and grammatical function affect pronoun production. *ExLing 2020* 113. https://doi.org/ExLing-2020/11/0028/000443

Mann, William C & Thompson, Sandra A. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8(3). 243–281. https://doi.org/10.1515/text.1.1988.8.3.243

Marcus, Mitchell P. & Santorini, Beatrice & Marcinkiewicz, Mary Ann. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2). 313–330. https://aclanthology.org/J93-2004.

Mayol, Laia. 2018. Asymmetries between interpretation and production in Catalan pronouns. *Dialogue & Discourse* 9(2). 1–34. https://doi.org/10.5087/dad.2018.201

Medina Fetterman, Ana M & Vazquez, Natasha N & Arnold, Jennifer E. 2022. The effects of semantic role predictability on the production of overt pronouns in Spanish. *Journal of Psycholinguistic Research* 1–26. https://doi.org/10.1007/s10936-021-09832-w

Patterson, Clare & Schumacher, Petra B & Nicenboim, Bruno & Hagen, Johannes & Kehler, Andrew. 2022. A Bayesian approach to German personal and demonstrative pronouns. *Frontiers in Psychology* 12. 6296. https://doi.org/10.3389/fpsyg.2021.672927

Poesio, Massimo & Artstein, Ron & Uryupina, Olga & Rodriguez, Kepa & Delogu, Francesca & Bristot, Antonella & Hitzeman, Janet. 2013. The ARRAU Corpus of Anaphoric Information. *Linguistic Data Consortium* https://doi.org/10.35111/y3mr-he10

R Core Team. 2021. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

Rohde, Hannah. 2019. Pronouns. In *The Oxford Handbook of Experimental Semantics and Pragmatics*, Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198791768.013.21

Rohde, Hannah & Kehler, Andrew. 2014. Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience* 29(8). 912–927. https://doi.org/10.1080/01690965.2013.854918

Rosa, Elise C. 2015. *Semantic role predictability affects referential form*. Chapel Hill: The University of North Carolina at Chapel Hill dissertation.

Rosa, Elise C & Arnold, Jennifer E. 2017. Predictability affects production: Thematic roles can affect reference form selection. *Journal of Memory and Language* 94. 43–60. https://doi.org/10.1016/j.jml.2016.07.007

Schad, Daniel J & Nicenboim, Bruno & Bürkner, Paul-Christian & Betancourt, Michael & Vasishth, Shravan. 2022. Workflow techniques for the robust use of bayes factors. *Psychological Methods* 28(6). 1404–1426. https://doi.org/https://doi.org/10.1037/met0000472

Schutze, Hinrich & Manning, Christopher D & Raghavan, Prabhakar. 2008. *Introduction to information retrieval*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Stevenson, Rosemary J & Crawley, Rosalind A & Kleinman, David. 1994. Thematic roles, focus and the representation of events. *Language and Cognitive Processes* 9(4). 519–

548. https://doi.org/10.1080/01690969408402130

Tily, Harry & Piantadosi, Steven. 2009. Refer efficiently: Use less informative expressions for more predictable meanings. In *Proceedings of the workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference*.

Verhagen, Véronique & Mos, Maria & Backus, Ad & Schilperoord, Joost. 2018. Predictive language processing revealing usage-based variation. *Language and cognition* 10(2). 329–373. https://doi.org/10.1017/langcog.2018.4

Weatherford, Kathryn C & Arnold, Jennifer E. 2021. Semantic predictability of implicit causality can affect referential form choice. *Cognition* 214. 104759. https://doi.org/10.1016/j.cognition.2021.104759

Webber, Bonnie & Prasad, Rashmi & Lee, Alan & Joshi, Aravind. 2019. The Penn Discourse Treebank 3.0 annotation manual. Linguistic Data Consortium, Philadelphia, PA.

Weischedel, Ralph & Palmer, Martha & Marcus, Mitchell & Hovy, Eduard & Pradhan, Sameer & Ramshaw, Lance & Xue, Nianwen & Taylor, Ann & Kaufman, Jeff & Franchini, Michelle & El-Bachouti, Mohammed & Belvin, Robert & Houston, Ann. 2013. Ontonotes release 5.0. https://doi.org/10.35111/xmhb-2b84. Linguistic Data Consortium, Philadelphia, PA

Zerkle, Sandra A & Arnold, Jennifer E. 2019. Does pre-planning explain why predictability affects reference production? *Dialogue & Discourse* 10(2). 34–55. https://doi.org/10.5087/dad.2019.202

Zhan, Meilin & Levy, Roger & Kehler, Andrew. 2020. Pronoun interpretation in Mandarin Chinese follows principles of Bayesian inference. *PLOS ONE* 15(8). e0237012. https://doi.org/10.1371/journal.pone.0237012