Supplementary file

A More information in the corpus analyses with verb types

This section presents our corpus analyses with transfer-of-possession verbs and implicit causality verbs. These particular linguistic elements have been widely employed in prior psycholinguistic research, therefore providing an intuitive basis for any attempt at replicating the findings in this field using corpus texts. However, we were unable to replicate the next-mention biases induced by these verbs as reported in previous studies. Despite this, we have detailed our methodology and findings in the hope that they will serve as valuable information to inform future research endeavors.

A.1 Analysis 1: transfer-of-possession verbs

We first examine the frequency of different thematic roles being the next mention and pronoun production in transfer-of-possession scenarios while controlling for grammatical roles. This aims at answering the following two questions: 1) How do transfer-of-possession verbs influence next-mention frequency in corpus texts? Are transfer-of-possession contexts more likely to continue with the Goal referent, i.e. are there more *goal* continuations in goal-source contexts than *source* continuations in source-goal contexts, as shown in previous studies? 2) If transfer-of-possession contexts more frequently continue with the Goal, are pronouns produced more often for the Goal in goal-source contexts than for the Source in source-goal contexts? Strong Bayes predicts uniform pronoun production rates. Alternatively, according to the Expectancy Hypothesis, verb semantics are predicted to influence not only next-mention biases but also pronoun production biases, leading to a higher pronominalization rate for Goal re-mentions in goal-source contexts than for Source re-mentions in source-goal contexts.

A.2 Method

To automatically extract corpus contexts that resemble the stimuli designed for story continuation tasks in psycholinguistic research, we first defined sets of source-goal verbs and goal-source verbs, identical to the ones that Arnold (2001) used in her study, listed in Table 16.

Type	Verbs
source-goal verbs	bring, give, hand, loan, offer, pass, pay, rent, sell, send, show, teach, tell, throw, toss
goal-source verbs	accept, borrow, buy, catch, get, grab, hear, inherit, learn, purchase, receive, rent, snatch, take

Table 16: Transfer-of-possession verbs used for context extraction.

Sentences containing a verb in either of these two lists as the main verb and three arguments (source, goal, and theme) were selected. Following Arnold (2001), we excluded sentences in which source-goal verbs are used in double-object constructions (*Anna gave Mary this book*) in order to maintain the consistency with goal-source verbs, in which the only possible construction for mentioning the Source is a prepositional phrase (*Mary received this book from Anna*). Source and Goal arguments in each sentence were then iden-

tified using the annotation of predicate-argument structure (see Table 4 for an example of annotations). After that, the first semantic argument annotated immediately following the transfer-of-possession construction was identified as the next mention. Most of the time, it is the grammatical subject of the following clause. Finally, we used coreference annotation to check which referent the next mention refers to, the Goal, the Source, or other referents, as illustrated in Figure 9.

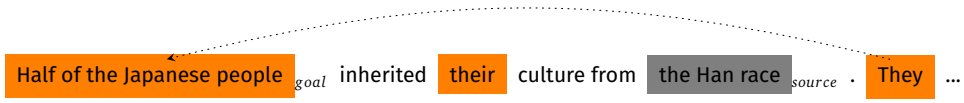

Figure 9: A context which continues with the Goal referent. Mentions marked with the same color refer to the same entity. They are in the same coreference chain and annotated with the same reference ID.

A.3 Results

For each verb type, we counted the number of segments where the next mention corefers with the Source antecedent, the Goal antecedent, the Theme antecedent, and other referents respectively. As to the last category, we consider all referents that have not been mentioned in the preceding clause (either new referents, or referents from earlier discourse). Following Arnold (2001), we compared the frequency of continuations referring back to the Goal and the Source with the grammatical function of the Goal/Source being controlled for, given a possible interaction between effects of grammatical functions and effects of thematic roles.

Figure 10 presents the percentage of continuations with Goal and Source antecedents, separately for when the antecedent is the subject and when it is the object (the raw number of samples for each coreference type is presented in Table 17).

antecedent	source-goal verbs (give)	goal-source verbs (receive)
source	98	9
goal	27	38
theme	44	8
other	174	112
Total	343	167

Table 17: Number of coreference samples automatically retrieved for transfer-of-possession contexts.

As we can see in Figure 10, for subject antecedents, Sources were mentioned more frequently than Goals, while the opposite was found for object antecedents. Overall, a chi-squared test shows that these differences were not significant and that Goal referents were mentioned as frequently as Source referents in continuations (subject antecedents: $X^2(1) = 1.66, p = .2$; object antecedents: $X^2(1) = .71, p = .4$).

Figure 11 presents the pronominalization rates. The analysis shows that there is no interaction between thematic role and pronominalization rate (subject antecedents: $X^2(1) = 1.23, p = .27$; object antecedents: $p = .47$). Though there were as many pronouns

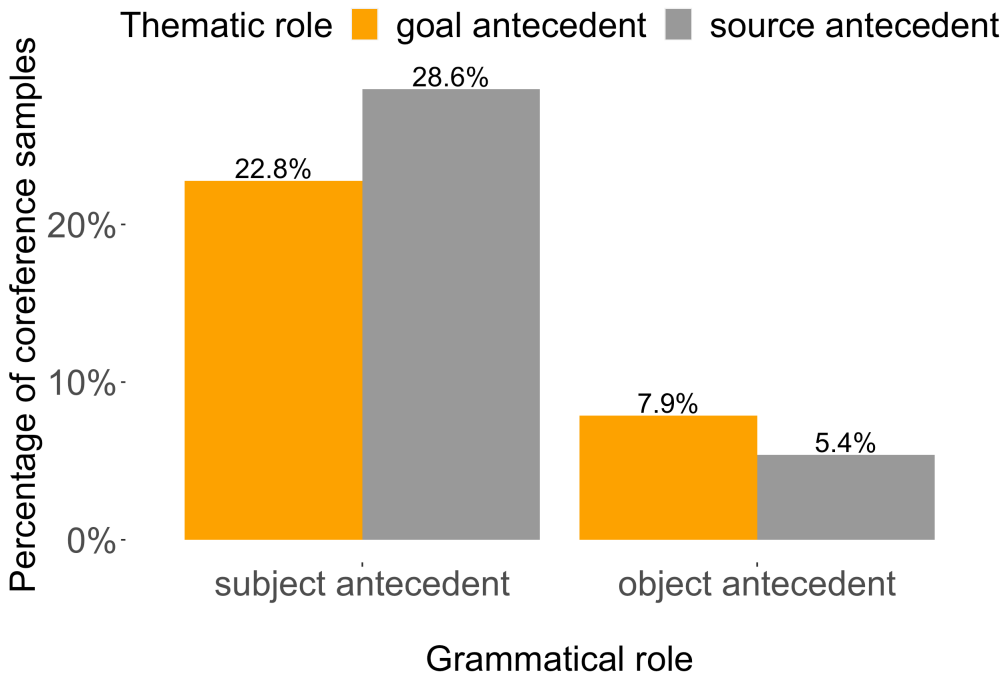

Figure 10: Percentage of Goal and Source continuations in two transfer-of-possession contexts.

produced to refer to the Source antecedents as to the Goal antecedents, evidence was insufficient for us to conclude that verb semantics does not affect pronoun production, given the lack of prior evidence showing one thematic role is more predictable than the other.

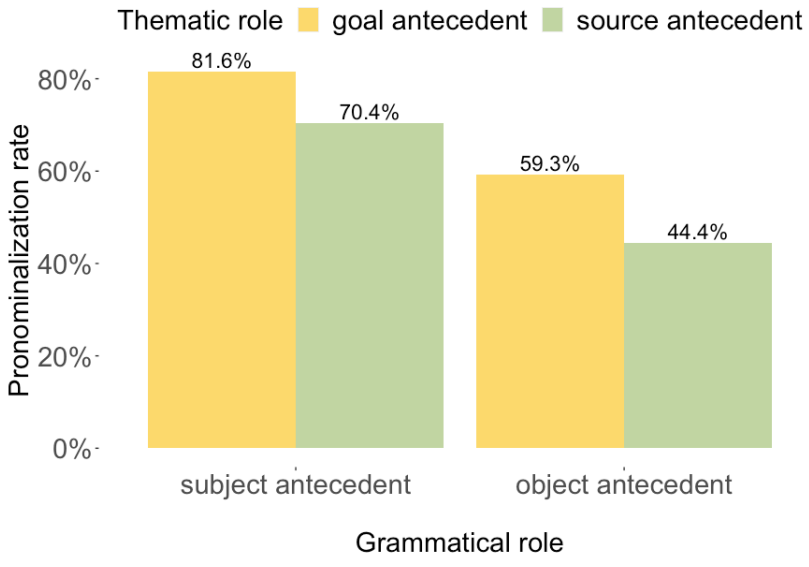

Figure 11: Pronominalization rate in each category for transfer-of-possession verbs.

Although these results seem to contradict previous findings that Goal referents are more frequently re-mentioned in transfer-of-possession contexts, our samples may not be sufficiently comparable to the designed material in terms of the common verb sense used in contexts.

Given that the corpus was annotated with semantic information, we attempted to reduce the effect of noise by specifying verb senses and restricting referents to personal pronouns as well as noun phrases denoting people, nationality, religious or political groups, organizations, countries, cities, or states. Nevertheless, this resulted in a sample size that was too small for further analysis.[19]

To sum up, in the analysis with transfer-of-possession verbs we did not obtain sufficient evidence to reach conclusions. Transfer-of-possession verbs in our corpus texts (especially news articles) more often depict abstract transfers, e.g., (19), rather than concrete ones as in the items of psycholinguistic experiments, e.g. (4). When used in other senses, transfer-of-possession verbs do not necessarily elicit a focus on the Goal.[20]

(19) The central government always gave strong backing to the special region's government and the compatriots of HK. We believed that …

A.4 Analysis 2: implicit causality verbs

This section presents our analysis with implicit causality verbs. We ask similar questions as in the previous analyses with transfer-of-possession verbs: 1) How do implicit causality verbs influence referent re-mention rates in corpus texts? Are subject-biased verbs (e.g., *scare*, *surprise*) more likely to continue with the subject, and object-biased verbs (e.g., *admire*, *dislike*), to the object, as shown in previous psycholinguistic research? 2) If the frequency patterns are congruent with that found in psycholinguistic research, are pronouns produced more often when referring back to the subject in subject-biased contexts than the subject in object-biased contexts?

A.4.1 Method

This analysis followed a similar methodology as in that with transfer-of-possession verbs. For the sake of simplicity, we will broadly characterize the more expected / predictable referent in all types of implicit causality verbs as the *implicit cause*. We attempted to extract contexts that resemble the experimental material in psycholinguistic research: sentences containing an implicit causality verb as the main verb and exactly two arguments. A set of object-biased and subject-biased verbs were selected from a corpus of implicit causality verbs (Ferstl et al. 2011).[21] All implicit causality verbs used in this extraction are listed in Appendix A.5. The two arguments and the next mention were then identified in a similar manner as described above for the extraction of transfer-of-possession contexts. Finally, the coreference chain of the next mention was compared against that of the two antecedents.

A.4.2 Results

We distinguish between contexts where the next mention corefers with the subject antecedent, the object antecedent, and other referents, in a similar manner as described

[19] This is not surprising given that OntoNotes is only partially annotated with named entities (mostly for news articles and for commonly-known entities). Many names in other genres such as narrative texts, and telephone conversations are not annotated.

[20] We tried to filter by verb sense, but then obtained too few samples for analysis.

[21] Ferstl et al. (2011) provide implicit causality bias scores of 300 English verbs on the basis of a sentence completion study in which participants were asked to add explicit explanations to fragments such as *John liked Mary because…*. We included verbs with either a subject bias or an object bias score larger than 65 (full score is 100).

previously for the analysis with transfer-of-possession verbs (Section A.3). Figure 12 shows the percentage of references to the subject and object antecedents (the raw numbers are presented in Table 18).

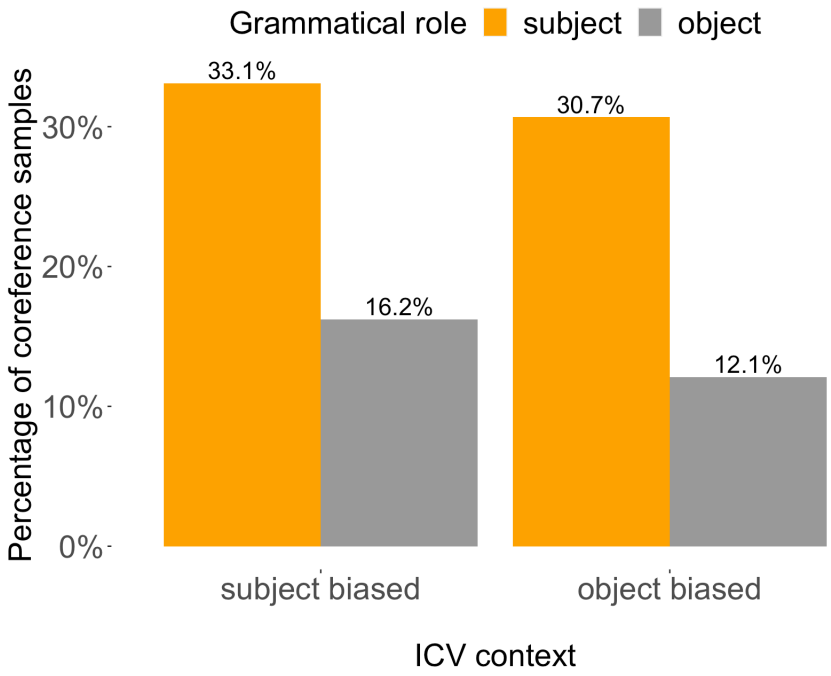

Figure 12: Percentage of continuations in subject-biased and object-biased implicit causality contexts.

	subject-biased verbs (surprise)	object-biased verbs (admire)
subject coref	49	104
object coref	24	41
total (including other coref)	148	339

Table 18: Number of coreference samples automatically retrieved for implicit causality verbs.

While previous psycholinguistic studies have shown that subject-biased verbs are biased towards the grammatical subject, and object-biased verbs, the grammatical object, we found a larger proportion of subject continuations relative to object ones in both. This is different from the findings in previous studies. We thus failed to replicate the contrasting likelihoods of next mention.

We present pronominalization rates in Figure 13 for extra information. In Figure 13, object antecedents in object-biased contexts and subject antecedents in subject-biased contexts were both coded as *implicit cause*, and the other argument in the context, in turn, *non implicit cause*. The analyses show that there was no difference between the amount of pronouns produced for *implicit cause* and that for *non implicit cause* (subject antecedents: $X^2(1) < 0$, $p = 1$; object antecedents: $X^2(1) = .08, p = .77$).

Like in transfer-of-possession scenarios, we again did not manage to replicate the biased patterns found in psycholinguistic studies. This is presumably due to the difference

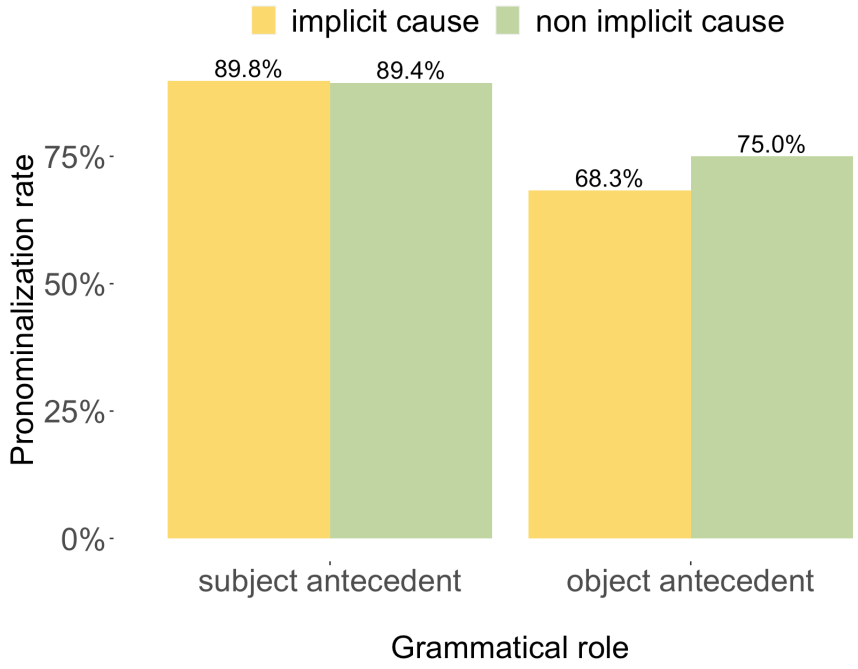

Figure 13: Pronominalization rate in each category for subject-biased and object-biased implicit causality verb contexts.

between naturally-occurring language in our corpus and controlled language in designed stimuli.

We did not try to further restrict referents, since this would probably lead to insufficient samples, especially for subject-biased verbs.

For implicit causality scenarios, subject-biased verbs in the corpus are more commonly used as predicates (e.g. *Anna was surprised that Mary* …), rather than in active verbal constructions (e.g. *Mary surprised Anna because* …) which are the ones used in psycholinguistic experiments. In addition, as the implicit cause is only more likely to be re-mentioned in Explanation (Kehler et al. 2008), most of the psycholinguistic studies on implicit causality verbs elicit continuations using connectives like *because*. However, we find that in the corpus, the cause is very often explained using prepositional phrases, as in (20). Contexts of this kind do not necessarily exhibit the observed pattern in Explanation because the noun phrase which is labeled as *next mention* comes after the cause has already been explained by a prepositional phrase.

(20) Russian Foreign Minister Igor Ivanov congratulated Kostunica on his election victory. He also gave him a letter from Russian President Vladimir Putin.

Yet another issue is the animacy of arguments. In transfer-of-possession contexts in the corpus, source-goal verbs are very likely to be used with an inanimate endpoint such as the location in (21), which reduces the probability of continuing with Goal referents. Implicit causality verbs have an analogous problem (see the object *the broader selection* in (22)). This makes it difficult to control for the topicality of arguments.

(21) The men brought their boats to the shore. They left …

(22) Jeanene Page, of North Salt Lake City, Utah, likes the broader selection. She wants something big …

To sum up, in the analyses with verb types we did not obtain sufficient evidence to reach conclusions.

A.5 List of implicit causality verbs

Subject-biased verbs

agitate amaze amuse anger annoy antagonize apologize appal attract betray bore bug call captivate charm concern confess daunt delight disappoint echo enrage enthral entice entrance exasperate excite fascinate frighten frustrate gladden infuriate inspire intimidate intrigue irritate lie madden mesmerise peeve please provoke repel repulse revolt scar sicken telephone trail trouble unnerve upset worry wow

Object-biased verbs

admire adore applaud appreciate calm carry celebrate comfort commend congratulate console correct counsel despise detest dislike distrust dread employ envy fancy favour fear feed guide hat idolize laugh like loathe love mourn notice penalize pick pity praise prize punish resent respect reward scold spank sue thank treasure value

B More information in the analyses with discourse relations

B.1 Raw sample counts of different coreference types

	Narration	Result	Contrast
subject coref	376	417	771
non-subject coref	180	213	455
other coref	429	700	1346
total	985	1330	2572

Table 19: OntoNotes: Counts of samples in each coreference type by discourse relation.

	Narration	Result	Contrast
subject coref	30	27	65
non-subject coref	8	17	35
other coref	33	104	249
total	71	148	349

Table 20: RST-DT: Counts of samples in each coreference type by discourse relation.

B.2 Robustness test of pronoun production analysis

For the sample from OntoNotes, 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.[22] This factor could potentially lead to varying lev-

[22] Given the limited number of samples with pronominal antecedents in RST-DT, we did not conduct further analysis to explore their potential influences on our results. Specifically, pronominal antecedents accounted for only 4% of all the extracted samples (25 out of 568; 5 in Narration, 14 in Contrast, and the other 6 in

	subject coreference		*non-subject coreference*		Total
	NM≠PRO	NM=PRO	NM≠PRO	NM=PRO	(incl. other coref)
Narration	7	23	7	1	71
Result	10	17	11	6	148
Contrast	13	52	28	7	349

Table 21: Raw pronoun data for coreference samples in RST-DT. PRO stands for pronoun. NM is abbreviation for next-mention.

els of topicality among the referents across different relations.[23] Furthermore, first- and second-person pronouns, such as *I* and *you*, inherently refer to the speaker or the addressee within the context of the utterance due to their deictic nature. As a result, referential choices other than pronouns are essentially eliminated. This differs from other referents, where speakers have the option to choose either a pronoun or a more explicit referring expression. Thus, a difference in the referential form of antecedent and the types of pronouns that appear with each discourse relation could invalidate the general results.

This additional test focuses on *subject* coreference contexts and applies another mixed-effect logistic regression model. We included discourse relation types as fixed effects, and incorporated the referential form of the subject antecedent (categorized into three levels: first- or second-person pronouns, other pronouns, and non-pronouns) as fixed effects. Random intercepts for document ID were included, as before. The results do not change. As displayed in Table 22, the rates of pronoun production do not exhibit variations across discourse relations, even after accounting for the influence of the antecedent's form. Furthermore, consistent with our expectations, we found that when the antecedent is expressed as a first- or second-person pronoun, there is a significantly higher likelihood that the re-mention will also be a pronoun. In contrast, when the antecedent is in a non-pronominal form, the likelihood of the next mention being a pronoun decreases.

Effects		Estimate	SE	z	*p*
Intercept		1.55	0.18	8.61	
discourse relation	Narration	−0.12	0.21	−0.60	0.55
	Result	−0.15	0.20	−0.75	0.45
antecedent type	1st, 2nd person pronoun	1.31	0.22	6.01	<0.001
	other pronoun	0.39	0.19	2.03	0.04

Table 22: Pronominalization of subject re-mentions in OntoNotes: mixed-effects logistic regression model with the next mention being a pronoun as the dependent measure.

Result). Therefore, their impact on our findings is deemed negligible and was not considered in the analysis of RST-DT.

[23] This is not an issue in psycholinguistic experiments, which typically introduce antecedents using names or full noun phrases.

B.3 Mapping between the original taxonomy of RST-DT and our categorization

Relation in OntoNotes	*Relation in RST-DT* (coarse-grained inventory)	*Relations in RST-DT* (fine-grained inventory)
Narration	Temporal	Temporal-before, Temporal-after, Temporal-same-time, Sequence, Inverted-sequence
Contrast	Contrast	Contrast, Concession, Antithesis
Result	Cause Explanation	Cause, Result, Consequence Evidence, Explanation-argumentative, Reason

Table 23: Relations in RST-DT that we deemed equivalent to those in OntoNotes. Note that the Result relation from OntoNotes distributes over both Cause and Explanation in RST-DT, due to the annotation decisions in RST-DT.

B.4 Extraction of discourse relations in the RST-DT corpus

In the RST-DT corpus, text segments are categorized according to their informational importance: a *nucleus* (N) represents the most essential piece of information in the relation, and a *satellite* (S) indicates supporting or background information (see Figure 14 for examples).

It is noteworthy that the assignment of nuclearity is determined by the semantic relevance of the information each units conveys, and therefore two syntactically equivalent text spans can be annotated with distinct discourse relations. For instance, while example (23) is annotated as a *Result* relation, example (24) is annotated as *Cause*, even though both are composed of a main clause followed by a subordinate clause that explains the cause of the event in the main clause. Therefore, the Result relation that we checked in OntoNotes can be extracted in RST-DT by specifying the structure to be *nucleus* + *satellite* in Cause or *satellite* + *nucleus* in Result. The extraction for the other two relations is more straightforward. In RST-DT, contexts of Narration and Contrast are mostly annotated as multinuclear relations (*nucleus* + *nucleus*), for which there is no directionality, as the two constituents are equally important. Contexts for these two relations are therefore directly extracted by specifying the name of relation.

(23) Result: [that next month's data isn't likely to be much better,]$_N$ [because it will be distorted by San Francisco's earthquake.]$_S$

(24) Cause: [Now this remarkable economic growth seems to be coming to an end]$_S$ [because the government has not converted itself into a modern, democratic, "developed nation" mode of operation.]$_N$

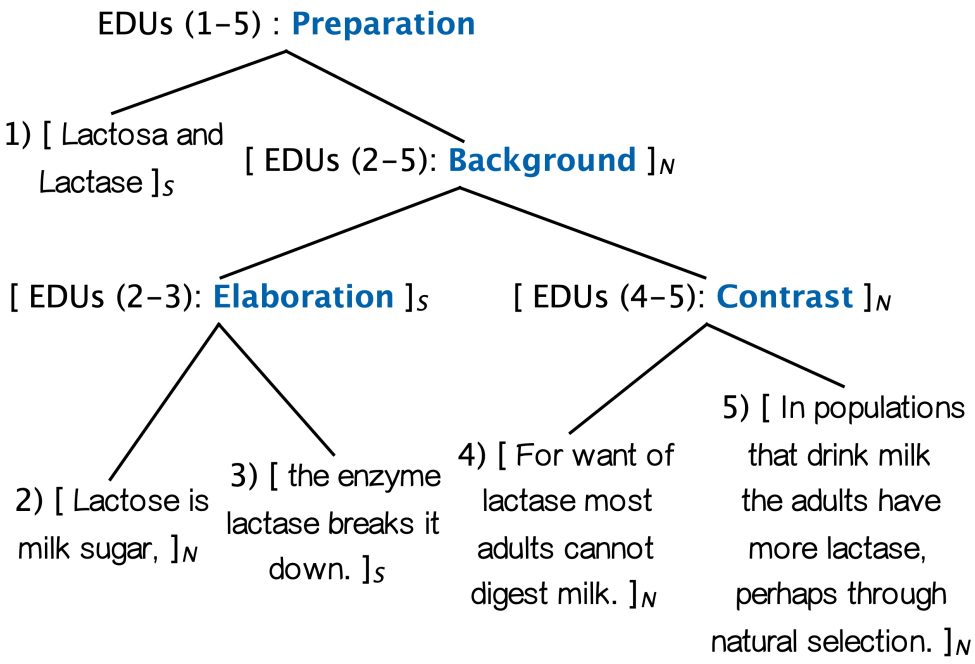

Figure 14: Graphical representation of an RST analysis, with nucleus/satellite annotated.

C Analysis results excluding participants with low variation in referring expressions

Following Rosa & Arnold (2017), we applied an exclusion criterion, excluding participants who produced fewer than 2 non-pronouns in the bare-prompt condition (all participants produced at least 2 pronouns). This led to the exclusion of 22 participants out of the initial pool of 200 participants. Our results and findings remain consistent both before and after this exclusion, as demonstrated below.

C.0.0.1 Next-mention biases

Our findings related to next-mention biases remain consistent with the results presented in the main text and support Hypothesis 1: subject referents are more predictable in Narration than in Contrast and Result. The summary of our mixed-logit model can be found in Table 24.

Fixed effects	Estimate	SE	Z	*p*
Intercept	0.51	0.25	2.00	
Narration	**0.66**	**0.14**	**4.75**	**<0.001**
Result	**−0.34**	**0.16**	**−2.12**	**0.03**

Table 24: Summary of logit mixed effect models of next mention with a fixed effect for the 3-level discourse relation type, excluding participants who produced fewer than 2 non-pronouns.

C.0.0.2 Pronoun production biases

As for Hypothesis 2, our analysis continues to yield no evidence of an effect of Relation Type on pronominalization, even after excluding participants with low variation in their referring expressions. This aligns with the findings presented in our main text. The results of the full Relation Type × Grammatical Role model are presented in Table 25.

	Estimate	Est.Error	95% CI
Intercept	0.81	0.30	[0.25, 1.41]
Narration	−0.03	0.23	[−0.47, 0.43]
Result	0.16	0.21	[−0.24, 0.58]
subject	**2.62**	**0.30**	**[2.07, 3.26]**
Narration:subject	0.32	0.26	[−0.17, 0.85]
Result:subject	-0.24	0.23	[−0.70, 0.22]

Table 25: Summary of logit mixed effect models of pronoun production (with all predictors centered), excluding participants who produced fewer than 2 non-pronouns.

Furthermore, we compared two Bayesian models: Model 1 (H1) and Model 2 (H0).[24] Model 1 included the predictor *Relation Type* in addition to *Grammatical Role* and their interaction, while Model 2 excluded *Relation Type* and considered only *Grammatical Role*. The Bayes Factor in favor of Model 2 (H0) over Model 1 (H1) was estimated to be 0.00002, 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 (the null hypothesis or Strong Bayes) is the more appropriate choice.

C.0.0.3 Pronoun interpretation biases

We re-estimated the mixed-logit model for 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 26.

Fixed effects	Estimate	SE	Z	*p*
Intercept	1.60	0.25	6.50	
Narration	**0.87**	**0.12**	**7.49**	**<0.001**
Result	**−0.41**	**0.12**	**−3.31**	**<0.001**
pronoun prompt	**1.08**	**0.11**	**9.45**	**<0.001**
Narration:pronoun prompt	**0.22**	**0.08**	**2.70**	**0.007**
Result:pronoun prompt	−0.09	0.07	−1.22	0.22

Table 26: Summary of logit mixed effect models of next mention with the fully crossed factors of Relation Type × Prompt Type (with all predictors centered), excluding participants who produced fewer than 2 non-pronouns.

[24] Model 1 shares the same fixed effect items with the model presented in Table 25, which are Relation Type, Grammatical Role, and the interaction between them. We simplified the random effect structure to include only the random intercepts for participants and items to facilitate the estimation of Bayes Factors.

In summary, even after excluding 22 participants with limited variations in their referring expressions, the results and findings remain consistent with those presented in our main text.

D Comparison of pronoun interpretation models using Bayesian methods

In Section 2.5, we discussed how Patterson et al. (2022) advanced model evaluation using Bayesian methods, which generate distributions of potential values rather than the point estimates utilized in the approaches of Rohde & Kehler (2014), Zhan et al. (2020), and our main text analysis.

We applied this Bayesian approach to reevaluate the three pronoun interpretation models. This not only complements our evaluation but also allows us to compare the outcomes derived from both the metrics and Bayesian methods. Our adaptation of the Patterson et al. (2022) method involved minor modifications to the model parameters to suit our experimental design. For a comprehensive explanation of the method, please refer to Patterson et al. (2022). The following sections will briefly introduce the method and then present our findings.

D.1 Method

Following Patterson et al. (2022), we use a Bayesian data analysis approach implemented in the probabilistic programming language *Stan* in R for the data analysis and modeling. Regularizing priors are used in all our models, which are minimally informative and have the objective of yielding stable inferences. We fit the models with four chains and 4000 iterations each, of which 1000 iterations were the burn-in or warm-up phase. In order to assess convergence, we verify that there are no divergent transitions.

D.1.0.1 Expectancy Model

The Expectancy Model assumes that the probability of referring to NP1 for pronoun prompt data is determined by the prior probability of NP1 ($P(referent = NP1)$). The model assumes that this prior can be estimated by the bare prompt data.

The Expectancy Model was build in the following way and its parameters were estimated using only the bare prompt data:

(VIII)

$$\begin{aligned}
\eta_i &= \alpha_{NP1} + u_{NP1}[subj_bare[i]] + w_{NP1}[item_bare[i]] \\
&\quad + rtype1[i] \cdot (\beta_{rtype1} + u_{rtype1}[subj_bare[i]]) \\
&\quad + rtype2[i] \cdot (\beta_{rtype2} + u_{rtype2}[subj_bare[i]]) \\
P(NP1|\ldots) &= P(referent = NP1 \mid item_bare[i], subj_bare[i], rtype1_i, rtype2_i) = logit^{-1}(\eta_i) \\
NP1_i &\sim Bernoulli(P(NP1|\ldots))
\end{aligned}$$

where $NP1$ is 1 if the referent is NP1, 0 if the referent is NP2, i indicates the observation of the bare prompt data, $rtype1$ and $rtype2$ are two vectors that map between observations and the corresponding relation type. In $rtype1$, Narration is coded with 1, Contrast coded with 0, and Result coded with -1. In $rtype2$, Narration is coded with 0, Contrast coded with 1, or Result coded with -1. $subj\_bare$ and $item\_bare$ are vectors that indicate the mapping between observations, and subjects and items respectively, and u

and w are the by-subject and by-items adjustments (or *random effects*). The three dots (...) symbolize all the information that the model is taking into account to estimate the probability of producing NP1 as a referent: the characteristics of the stimuli (i.e., intercept, beta, and by-item adjustments) and of the subject performing the bare prompt task (i.e., by-subject adjustments).

The parameters estimated with the bare prompt data were used to generate predictions for the pronoun prompt data in the following way:

(IX)

$$\begin{aligned}
\eta_n =& \alpha_{NP1} + u_{NP1}[subj_pron[n]] + w_{NP1}[item_pron[n]] \\
&+ rtype1[n] \cdot (\beta_{rtype1} + u_{rtype1}[subj_pron[n]]) \\
&+ rtype2[n] \cdot (\beta_{rtype2} + u_{rtype2}[subj_pron[n]]) \\
P(NP1 \mid \ldots) =& P(referent = NP1 \mid item_pron[n], subj_pron[n], rtype1_n, rtype2_n) = logit^{-1}(\eta_n) \\
pred_{NP1_n} \sim& Bernoulli(P(NP1 \mid \ldots))
\end{aligned}$$

where n indicates the observation of the pronoun prompt data, $subj\_pron$ and $item\_pron$ are vectors that indicate the mapping between observations, and subjects and items respectively, and u and w are the by-subject and by-items adjustments. Similarly as before, the three dots (...) symbolize all the information that the model is taking into account generate the predictions: the characteristics of the stimuli (i.e., intercept, beta, and by-item adjustments) and of the subject performing the pronoun prompt task (i.e., by-subject adjustments).

D.1.0.2 Mirror Model

The Mirror Model assumes that the probability of referring to NP1 for pronoun prompt data is determined by the pronoun production likelihood ($P(pronoun|referent = NP1)$). The model assumes that this likelihood can be estimated from the bare prompt data.

The mirror model was build in the following way:

(X)

$$\begin{aligned}
\zeta_i =& \alpha_{pro} + u_{pro}[subj_bare[i]] + w_{pro}[item_bare[i]] \\
&+ rtype1[i] \cdot (\beta_{pro,rtype1} + u_{pro,rtype1}[subj_bare[i]]) \\
&+ rtype2[i] \cdot (\beta_{pro,rtype2} + u_{pro,rtype2}[subj_bare[i]]) \\
&+ ref_bare_i \cdot (\beta_{pro,ref} + u_{pro,ref}[subj_bare[i]] + w_{pro,ref}[item_bare[i]]) \\
&+ rtype1[i] \cdot ref_bare_i \cdot (\beta_{pro,int1} + u_{pro,int1}[subj_bare[i]] + w_{pro,int1}[item_bare[i]]) \\
&+ rtype2[i] \cdot ref_bare_i \cdot (\beta_{pro,int2} + u_{pro,int2}[subj_bare[i]] + w_{pro,int2}[item_bare[i]]) \\
P(pro \mid \ldots) =& logit^{-1}(\zeta_i) \\
pron_i \sim& Bernoulli(P(pro \mid \ldots))
\end{aligned}$$

where $pron$ is 1 if the bare completion includes a pronoun, 0 if it includes a non-pronoun, i indicates the observation of the bare prompt data, $rtype1$ and $rtype2$ are two vectors that map between observations and the corresponding relation type. In $rtype1$, Narration is coded with 1, Contrast coded with 0, and Result coded with -1. In $rtype2$, Narration is coded with 0, Contrast coded with 1, or Result coded with -1. ref_bare indicates whether the referent of the completion is NP1 (coded with 1) or NP2 (coded with -1), and as for the Expectancy Model, $subj\_bare$ and $item\_bare$ are vectors that indicate the mapping between observations, and subjects and items respectively, and u and w are the by-subject and by-items adjustments.

The parameters estimated with the bare prompt data were used to generate predictions for each observation n of the pronoun prompt data in the following way.

First the likelihood of each referent is calculated:

(XI)

$$\begin{aligned}
P(pro \mid NP1, \ldots) = {} & \alpha_{pro} + u_{pro}[subj_pron[n]] + w_{pro}[item_pron[n]] \\
& + rtype1[n] \cdot (\beta_{pro,rtype1} + u_{pro,rtype1}[subj_pron[n]]) \\
& + rtype2[n] \cdot (\beta_{pro,rtype2} + u_{pro,rtype2}[subj_pron[n]]) \\
& + (\beta_{pro,ref} + u_{pro,ref}[subj_pron[n]] + w_{pro,ref}[item_pron[n]]) \\
& + rtype1[n] \cdot (\beta_{pro,int1} + u_{pro,int1}[subj_pron[n]] + w_{pro,int1}[item_pron[n]]) \\
& + rtype2[n] \cdot (\beta_{pro,int2} + u_{pro,int2}[subj_pron[n]] + w_{pro,int2}[item_pron[n]]) \\
P(pro \mid NP2, \ldots) = {} & \alpha_{pro} + u_{pro}[subj_pron[n]] + w_{pro}[item_pron[n]] \\
& + rtype1[n] \cdot (\beta_{pro,rtype1} + u_{pro,rtype1}[subj_pron[n]]) \\
& + rtype2[n] \cdot (\beta_{pro,rtype2} + u_{pro,rtype2}[subj_pron[n]]) \\
& + (-1)(\beta_{pro,ref} + u_{pro,ref}[subj_pron[n]] + w_{pro,ref}[item_pron[n]]) \\
& + rtype1[n] \cdot (-1) \cdot (\beta_{pro,int1} + u_{pro,int1}[subj_pron[n]] + w_{pro,int1}[item_pron[n]]) \\
& + rtype2[n] \cdot (-1) \cdot (\beta_{pro,int2} + u_{pro,int2}[subj_pron[n]] + w_{pro,int2}[item_pron[n]])
\end{aligned}$$

Then the probability of the referent NP1 is calculated:

(XII)

$$P(NP1 \mid pro, \ldots) = \frac{P(pro \mid NP1, \ldots)}{P(pro \mid NP1, \ldots) + P(pro \mid NP2, \ldots)}$$

This probability is used to predict each observation n:

(XIII)

$$pred_{NP1_n} \sim Bernoulli(P(referent \mid pro, \ldots))$$

As before, the ... symbolize all the information that the model is taking into account generate the predictions: the characteristics of the stimuli (i.e., intercept, beta, and by-item adjustments) and of the subject performing the bare prompt task (i.e., by-subject adjustments).

D.1.0.3 Bayesian Model

The Bayesian Model assumes that the probability of referring to NP1 for pronoun prompt data is determined by its posterior distribution in the bare prompt data according to the Bayes's rule: the likelihood of NP1 ($P(pronoun|referent = NP1)$) is multiplied by the prior probability of NP1 ($P(referent = NP1)$) normalized to be a probability distribution by dividing it by the marginal probability distribution of the pronouns. The model assumes that this posterior can be estimated by the bare prompt data.

The parameters of the Bayesian Model are estimated using equations (VIII) from the Expectancy Model and (X) from the Mirror Model. In addition, the by-participants and by-items adjustments from both (VIII) and (X) are used.

These parameters estimated with the bare prompt data were used to generate predictions for each observation n of the pronoun prompt data in the following way.

We calculate the prior $P(NP1)$ based on equation (IX) and the likelihood $P(pro \mid NP1)$ based on equations (XI). With these we calculate $P(NP1 \mid pro)$.

The posterior probability of the referent NP1 is calculated conditional on a pronoun:

(XIV) $$P(NP1 \mid pro, \ldots) = \frac{P(pro \mid NP1)P(NP1)}{P(pro \mid NP1)P(NP1) + P(pro \mid NP2)(1 - P(NP1))}$$

This probability is used to predict each observation n:

(XV) $$pred_{NP1_n} \sim Bernoulli(P(NP1 \mid pro, \ldots))$$

D.2 Model comparison

We compare the models numerically using the expected log-predictive density (elpd) score of the models, with a higher score indicating better predictive accuracy for the held out pronoun-prompt data. The table 27 shows a clear advantage in predictive accuracy for the Bayesian model: When the difference between predictive density (*elpd_diff*) is larger than four, and the number of observations is larger than 100 then the normal approximation and the standard errors are quite reliable description of the uncertainty in the difference. As a rule of thumb, differences larger than 4 are considered enough to differentiate the predictive performance of the models. We also calculated the *weight* of the predictions of each model models by model averaging via stacking of predictive distributions. Stacking maximizes the potential elpd score by pulling the predictions of all the different models together. The values under the weight column represent the relative contribution of each model to the combined optimal model. In this case, the Bayesian model alone contributes almost 100% to the weighted predictions.

	elpd_diff	se_diff	elpd	se_elpd	weight
Bayesian	0	0	−999	33	0.94
Mirror	−205	21	−1204	33	0.03
Expectancy	−526	31	−1525	27	0.03

Table 27: Table: The table is ordered by the expected log-predictive density (elpd) score of the models, with a higher score indicating better predictive accuracy. The highest scored model is used as a baseline for the difference in elpd and the difference standard error (SE). The column weight represents the weights of the individual models that maximize the total elpd score of all the models.

In Table 28, we compare the Mirror and the Expectancy Models alone. It is clear that the Mirror model has a predictive performance superior to the Expectancy model.

	elpd_diff	se_diff	elpd	se_elpd	weight
Mirror	0	0	−1204	33	0.77
Expectancy	−321	42	-1525	27	0.23

Table 28: Table: The table is ordered by the expected log-predictive density (elpd) score of the models, with a higher score indicating better predictive accuracy. The highest scored model is used as a baseline for the difference in elpd and the difference standard error (SE). The column weight represents the weights of the individual models that maximize the total elpd score of all the models.

We show in Table 29, the expected log-predictive density (elpd) score of the models assessed for three subsets of the data, depending whether the relation type is Narration, Contrast, or Result.

Model	elpd_diff	se_diff	weight
Narration			
Bayesian	0.0	0.0	1.00
Mirror	−55.0	8.1	0.00
Expectancy	-185.0	13.5	0.00
Contrast			
Bayesian	0.0	0.0	0.81
Mirror	−51.0	12.7	0.15
Expectancy	-170.0	18.4	0.04
Result			
Bayesian	0.0	0.0	0.88
Mirror	−99.0	14.0	0.02
Expectancy	−171.0	21.0	0.10

Table 29: Difference in expected log-predictive density (elpd_diff) of the models assessed for three subsets of the data from the experimental work, depending on whether the relation type is Result, Narration, or Contrast.