Referent Predictability in Pronoun Production: Insights from a Multi-Methodological Exploration

Xixian Liao

TESI DOCTORAL UPF / Year 2023

THESIS SUPERVISOR
Gemma Boleda, Laia Mayol
Department of Translation and Language Sciences

Universitat
Pompeu Fabra
Barcelona

Acknowledgements

At the age of 14, I was introduced to the essence of what it means to be a teacher through the essay 师说 (“On Teaching”) by the Tang dynasty scholar 韩愈 (Han Yu). Ever since, I’ve nurtured the dream of becoming one myself. Over the past four years, numerous researchers have played pivotal roles in my journey. Their guidance, teachings, and insights have been invaluable.

> 师者，所以传道授业解惑也。– 韩愈《师说》
>
> “*A teacher is who could propagate the doctrine, impart professional knowledge, and resolve doubts* – On Teaching, Han Yu”

Foremost among them are my supervisors, Gemma Boleda and Laia Mayol. Their mentorship transcends the definition of a teacher as described by Han Yu. Our weekly meetings, shared coffees, lunches, and the moments of laughter and support will forever have a special place in my heart. They have not only guided my academic pursuits but also helped me discover my true desires and aspirations. I believe they’ve left a lasting impression on my heart, evoking emotions whenever I recall them in the years to come. In the words of a Catalan saying we share, I’ll continue to “anar fent” – keep moving forward.

I’d also like to express my appreciation to Hannah Rohde. While she wasn’t my official supervisor, she has been a steadfast presence since the defense of my research proposal in my first year of the Ph.D. program. From my research stay at the University of Edinburgh to the collaborative effort on our article this year, Hannah has been there. Her patient and vivid explanations, along with her poetic choice of words, never cease to amaze me. She has a unique ability to exude a calming influence, and her insights, prompt responses to my late-night emails, and assistance with my research questions have been invaluable.

In addition, Bonnie Webber has been an immense source of inspiration in the last two years of my Ph.D. I have fond memories of crossing paths with her on Edinburgh’s streets as I walked between the city center and Leith. She radiates boundless energy and a passion for research that makes time seem to stand still. Our weekly meetings in Edinburgh and virtual collaborations during my co-supervision of the master’s student, Yunfang Dong, with her in the last year of my Ph.D., have left an indelible mark on my academic journey. I can’t help but feel that something would be missing if I were to stop having meetings with her. Bonnie is always generous with praise and encouragement for junior researchers, and I’ll never forget her email that simply said, “Xixian - You are so wonderfully positive!” It brought happiness to my days for quite some time. She is undoubtedly one of the main reasons why my memories of the days spent in Edinburgh are always bathed in sunshine.

My heartfelt thanks also go to my collaborators, Laura Aina, Thomas Brochhagen, and Matthijs Westera. Their companionship and shared passion for research have been a constant source of support throughout these transformative years. Laura has surely not only been a collaborator but also a guiding presence and dear friend. We've shared many moments, both in work and in our personal lives, and yes, even some Taylor Swift songs. Working with Thomas, the statistics guru in our department, has been an honor. Our discussions on meta-analysis felt both intellectually stimulating and comforting, like a warm cup of tea on a chilly day.

I also wish to express my deep gratitude to the members of my doctoral committee evaluating this thesis: Jennifer Arnold, Andrew Kehler, Yufang Hou, Louise McNally, and Xavier Villalba. Their research has been a constant source of inspiration. I'm happy that you all accepted to be on board and really appreciate you devoting your time to my work.

> 学贵得师，亦贵得友。– 唐甄《潜书·讲学》
>
> "*In the pursuit of knowledge, a teacher is a treasure; so too is a friend.*
> – Hidden Books: On Learning, Tang Zhen"

As the philosopher and educator Tang Zhen noted during the late Ming and early Qing dynasties, not only is the guidance of teachers invaluable in our quest for knowledge, but the support and encouragement from friends also hold a special place.

Barcelona seems to possess a certain enchantment. It draws in not just brilliant researchers but also genuinely wonderful individuals. The thought of no longer being able to walk into the office, see Lucas Weber engrossed in work facing to the wall from as early as 8 am, hear the warm "Hola" in the morning, or respond to the familiar midday query "coffee?" invariably brings a tear to my eye. Countless happy hours on Friday night, celebrations, and weekend gatherings over these past four years with my colleagues in the COLT and GLiF groups at Universitat Pompeu Fabra have left an indelible mark. My heartfelt thanks go to Sara Amido, Marco Baroni, Marina Bolea Centelles, Jeanne Bruneau-Bongard, Sebastian Buchczyk, Nathanaël Carraz Rakotonirina, Emily Cheng, Roberto Dessì, Francesca Franzon, Eleonora Gualdoni, Corentin Kervadec, Matéo Mahaut, Guillermo Montaña Calverde, Dominika Slušná, and IonutTeodor Sorodoc.

And then, there's my cherished Chinese community: 黄紫 (Zi Huang), 梁嘉玲 (Jialing Liang), 郗晓彤 (Xiaotong Xi), 任能静 (Nengjing Ren), 张涵 (Han Zhang), and 王文元 (Wenyuan Wang). Together, we've traveled, surfed, snowboarded, shared drinks, and sought out the best Chinese restaurants in Barcelona. While each of us pursues our Ph.D. far from home and family, I find solace among these remarkable individuals.

The research presented in this thesis was made possible through the support of several grants and funding sources. These are the FI-AGAUG grant (2020FI-B00575), the grant PGC2018-094029-A-I00 from the Spanish State Research Agency (AEI) and European Regional Development Fund (FEDER), the grant PID2021-122779NB-I00 funded by Ministerio de Ciencia e Innovación(MCIN) / AEI / 10.13039 / 501100011033 and European Regional Development Fund - "A way of making Europe", funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154), as well as financial support from the Catalan government (SGR 2021 00470) and the Department of Translation and Language Sciences at Universitat Pompeu Fabra.

最深沉的感谢给我的父母。爸，妈，我爱你们！(My deepest gratitude goes to my parents. Dad, Mom, I love you!)

Finally, in my moments of doubt and contemplation, I found solace in the timeless words of Wang Wei from his evocative poem, "Zhongnan Mountain Retreat". I've included these lines not just as a personal reminder for when I revisit this work in the future, but also as a guiding light for those who may find themselves grappling with challenges in their own journeys:

> 行到水穷处, 坐看云起时。– 王维《终南别业》
>
> "*Where water ends, clouds rise.* – Zhongnan Mountain Retreat, Wang Wei"

May we always find the strength to push through, even when the path ahead seems uncertain！

廖茜娴 (Xixian Liao)
Barcelona, October 2023

Abstract

While it is known that speakers tend to use more reduced forms (e.g., by shortening or deleting segments) for words or phrases that are semantically predictable from context, it is contentious whether the same phenomena occurs at the referential level, with divergent findings reported in the literature. Specifically, it is unclear if more reduced referential forms, such as pronouns, are used more frequently for predictable referents. This thesis explores this question via a series of novel, primarily computational, approaches: analyses of richly annotated corpora, a corpus passage continuation task with human participants, derivation of predictability estimates from a neural network model, and a Bayesian meta-analysis. The findings from this thesis align more closely with the view that referent predictability does influence pronoun usage, albeit to a modest extent. Speakers are rational and efficient, choosing more reduced forms like pronouns for more predictable referents.

Resum

Tot i que és ben sabut que els parlants tendeixen a utilitzar formes més reduïdes (p. ex., escurçant o eliminant segments) per a paraules o sintagmes que són semànticament previsibles a partir del context, continua sent controvertit si succeeix el mateix fenomen al nivell referencial, amb resultats divergents a la literatura. Concretament, no està clar si formes referencials més reduïdes, com els pronoms, s'utilitzen més freqüentment per a referents previsibles. Aquesta tesi explora aquesta qüestió mitjançant nous mètodes, principalment computacionals: anàlisis de corpus extensament anotats, una tasca de continuació de fragments de corpus amb participants humans, derivació d'estimacions de previsibilitat a partir d'un model de xarxa neuronal, i una metaanàlisi bayesiana. Els resultats d'aquesta tesi s'alineen millor amb la visió que la previsibilitat del referent sí que influeix en l'ús de pronoms, encara que de forma modesta. Els parlants són racionals i eficients, i escullen formes més reduïdes com els pronoms per a referents més previsibles.

Resumen

Si bien es conocido que los hablantes tienden a usar formas más reducidas (por ejemplo, acortando o eliminando segmentos) para palabras o sintagmas que son semánticamente predecibles por el contexto, sigue siendo controvertido si ocurre el mismo fenómeno a nivel referencial, con resultados divergentes en la literatura. Específicamente, no está claro si se utilizan más frecuentemente formas referenciales más reducidas, como los pronombres, para referentes predecibles. Esta tesis explora esta cuestión mediante métodos novedosos, principalmente computacionales: análisis de corpus extensamente anotados, una tarea de continuación de fragmentos de corpus con participantes humanos, derivación de estimaciones de previsibilidad a partir de un modelo de red neuronal, y un metaanálisis Bayesiano. Los hallazgos de esta tesis se alinean mejor con la visión de que la previsibilidad del referente sí influye en el uso de pronombres, aunque en una medida modesta. Los hablantes son racionales y eficientes, eligiendo formas más reducidas como los pronombres para referentes más predecibles.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

When people process language, they use contextual information to make predictions about what will come next. For instance, before encountering the blank within the sentence: “She spread the warm peanut butter on the _____”, we have already formulated expectations about the probable subsequent word (for example, “bread”, in this case). This predictive nature of processing has received significant attention in recent research in Cognitive Science and Psycholinguistics (see, inter alia, Bubic et al. 2010; Clark 2013; Kuperberg and Jaeger 2016). The degree to which an addressee can anticipate what will come next based on contextual cues and prior knowledge is referred to as **predictability**. Work on language comprehension shows that the more predictable some linguistic input is, the faster and more accurately it is processed by addressees (see, a.o., Smith and Levy 2013; Staub 2015; Kuperberg and Jaeger 2016). Therefore, the role of predictability is clearly established in processing, but does predictability also guide language production?

In this dissertation, I focus on the role of predictability in language production, specifically in the production of referring expressions. This is crucial to understand the underlying mechanisms that facilitate coordination between speakers and addressees (e.g., Keysar et al., 2000; Pickering and Garrod, 2004; Brown-Schmidt and Tanenhaus, 2008): whether the mechanisms that speakers and addressees use are fully aligned, or whether it is possible that speakers ignore cues that addressees are sensitive to.

Over the last decades, the selection of referring expressions by speakers in discourse has been extensively investigated within the realm of language production. When referring to discourse entities, speakers can use a variety of expressions, such as proper names (“Angela Merkel”), descriptions (“the first female chancellor of Germany”), or more reduced expressions like pronouns (“she”). To what extent is this choice influenced by the speaker’s consideration of the addressee’s expectations? There is abundant evidence from other domains that speakers tend to use more reduced or attenuated forms for more predictable words or phrases (e.g., Lieberman, 1963; Jurafsky et al., 2001; Aylett and Turk, 2004; Bell et al., 2009; Piantadosi et al., 2012; Jaeger and Buz, 2017). For instance, the highly predictable "nine" in "A stitch in time saves ..." tended to exhibit shorter duration, reduced amplitude, and less precise articulation compared to the pronunciation of "nine" in "The number that you will hear is ...". Predictive processing frameworks suggest that these reductions enhance communicative efficiency by allowing for less speaker efforts without incurring into significant communicative cost (e.g., Jaeger and Levy, 2006).

Given this background, it is natural to assume that predictability will influence the production of referring expressions: For example, using pronouns is a more efficient way to refer to a previously mentioned entity that is already established in the discourse, especially when addressees can easily anticipate which entity the speaker is referring to. In this context, then, referent predictability is construed as the addressee's estimate of the likelihood that the speaker will mention a specific referent in the upcoming discourse. However, the numerous studies that have explored this question in the last two decades have yielded mixed results. Some studies found that speakers/writers are more likely to use pronouns when the referent is deemed to be more likely to be mentioned next, (e.g. Arnold, 2001; Tily and Piantadosi, 2009; Rosa and Arnold, 2017; Zerkle and Arnold, 2019; Lindemann et al., 2020; Konuk and von Heusinger, 2021; Weatherford and Arnold, 2021; Medina Fetterman et al., 2022; Hwang, 2023b), while others did not (e.g. Ferretti et al., 2009; Fukumura and Van Gompel, 2010; Rohde and Kehler, 2014; Rosa, 2015; Holler and Suckow, 2016; Modi et al., 2017; Mayol, 2018; Kehler and Rohde, 2019; Zhan et al., 2020; Frederiksen and Mayberry, 2022; Hwang et al., 2022; Kravtchenko, 2022; Lam and Hwang, 2022; Patterson et al., 2022; Hwang, 2023a).

This discrepancy in findings has given rise to a long-standing debate in the field; and to theories that make divergent predictions (see Section 2.2). Those who argue in favor of the role of referent predictability in pronoun production argue that speakers are more likely to use a pronoun when the referent is highly predictable (e.g., Arnold, 2001; Tily and Piantadosi, 2009), as pronouns provide more efficient means of referring to entities, signalling to addressees to retrieve the most accessible referent in memory. On the other hand, those who argue against the role of referent predictability in pronoun production suggest that other factors, such as grammatical and structural factors, play a more central role in pronoun production (e.g., Fukumura and Van Gompel, 2010; Rohde and Kehler, 2014).

Amidst this inconclusive picture of the existing literature, it's notable that these investigations mostly adhere to very similar psycholinguistic experiments, wherein participants are tasked with producing a continuation for given contexts. These contexts are often constructed to examine expectation biases on the next mention elicited by two specific verb types: transfer-of-possession verbs, such as “give” and “receive”, and implicit causality verbs, such as “surprise” and “admire”. Moreover, they are 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”. Importantly, such contexts are infrequently observed in corpora of natural language (see Chapter 3 and Appendix A).

The uniformity in the methodology and the limited representation of naturalistic scenarios highlight the need to explore alternative methods and use contexts that are more representative of naturally occurring language. This dissertation aims to offer new insights into the relationship between referent predictability and pronoun production via a multi-methodological exploration of English data. Through a series of novel, primarily computational, methods, we also hope to showcase how the rapid computational advancements over recent decades could benefit linguistic research.

1.1. Structure of the thesis

The rest of the dissertation is organized as follows:

Chapter 2: Background I provide an overview of factors that influence pronoun production and review models that posit different kinds of relationships between referent predictability and pronoun usage. I then introduce the methods commonly employed in previous research to operationalize and manipulate referent predictability. These approaches form the foundation upon which the research in this thesis builds its methodological framework.

Chapter 3: Corpus Analyses I automatically extract passages from two richly-annotated corpora developed in the computational linguistic research. Using predictability of referents estimated based on their next-mention frequency across corpus texts, I investigate how predictability influences pronoun usage in naturalistic corpus texts.

The content of this chapter is under review and partially based on the following publication:

- Xixian Liao (2022). Coherence-driven predictability and referential form: Evidence from English corpus data. *Proceedings of Sinn Und Bedeutung*, 26, 544–556.

Chapter 4: Corpus Passage Completion To obtain a more robust set of evidence supporting the observational findings in the corpus analyses, I use a set of extracted corpus passages as stimuli for a passage completion experiment with human participants. I additionally investigate how referent predictability influences pronoun interpretation and quantitatively compare three proposed models of pronoun interpretation.

At the time of writing, an article based on the content of this chapter is under review.

Chapter 5: Computational Modeling I use computational estimates of referent predictability from a neural network model and investigate the relationship between referent predictability and form of referring expression (both its syntactic type and length).

The content of this chapter is based on the following publication:

- Laura Aina, Xixian Liao, Gemma Boleda, Matthijs Westera (2021). Does referent predictability affect the choice of referential form? A computational approach using masked coreference resolution. In *Proceedings of the 2021 CoNLL Conference on Computational Natural Language Learning*.

Chapter 6: Meta-analysis I carry out a Bayesian meta-analysis, which covers 20 primary peer-reviewed studies, encompassing 26 samples across 8 languages. This is the first comprehensive synthesis of available evidence on the relationship between referent predictability and pronoun production.

At the time of writing, the content of this chapter is under review.

Chapter 7: Discussion I discuss the contributions of the thesis and explore potential avenues for future research, drawing upon insights from the various studies presented in the thesis.

Chapter 2

BACKGROUND

Language production is a complex process influenced by a multitude of factors. The decision to produce a pronoun instead of other referential forms is termed *pronoun production*, which has been a subject of extensive research. This chapter first reviews some of the most well-known contextual factors that influence pronoun production, which are outlined below. I then review contrasting proposals from the literature with respect to predictability and pronoun use. Following this, I will survey a range of tasks and materials that have been used to investigate the subject matter.

2.1. Factors influencing pronoun production

One crucial factor that affects pronoun production is grammatical function (e.g., Crawley et al., 1990; Brennan, 1995). Referents in subject position or with higher grammatical prominence e.g., “Brittany” in Ex. (1-a) and “Amanda” in Ex. (1-b), are more likely to be re-mentioned using pronouns, compared to other more oblique grammatical roles e.g., “Amanda” in Ex. (1-a) and “Brittany” in Ex. (1-b).

(1) a. Brittany admired Amanda.
b. Amanda amazed Brittany.

Information structural factors, such as topichood, are also known to affect pronoun production (e.g., Rohde and Kehler, 2014). Pronouns are often used for referents deemed central or topical within a discourse, presumably because these entities are cognitively more accessible. For example, pronouns are more frequently used to refer back to subjects in passive voice (e.g., “Brittany” in Ex. (2-a)) than those in active voice (e.g., “Amanda” in Ex. (2-b)), which are typically viewed as less topical than the former.

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

Beyond grammatical and structural factors, animacy likewise affects pronoun production (e.g., Fukumura and Van Gompel, 2011). Animate entities, such as people or animals, are more likely to be referred to by pronouns, probably due to their enhanced accessibility in memory and increased likelihood of being the focus of attention within discourse. For example, a speaker may be more inclined to use a pronoun for people like "The hikers" in Ex. (3-a) than for inanimate objects such as "The canoes" in Ex. (3-b).

(3) a. The hikers carried the canoes a long way downstream. Sometimes, _______________
    b. The canoes carried the hikers a long way downstream. Sometimes, _______________

Competition, i.e., the presence of multiple potential referents in the discourse, is another factor that plays a role in pronoun production (e.g., Arnold and Griffin, 2007). To avoid ambiguity, speakers might prefer a full noun phrase or name over a pronoun when several referents share similar properties e.g., Ex. 4.

(4) John and Mike went to the park. John played basketball, while Mike enjoyed the playground.

Additionally, the frequency of a referent's appearance in discourse influences its accessibility and pronoun production (e.g., Ariel, 1990). Referents mentioned more frequently are more likely to be refered to by pronouns. For example, in a text about a famous artist, the artist's name may often be substituted with a pronoun after multiple mentions. Relatedly, recency is another factor that has been shown to influence pronoun production (e.g., McCoy and Strube, 1999; Ariel, 2001). According to Arnold (2010), referents that have been mentioned recently are more accessible than those that have not. The more recent a referent's mention, the higher the probability of its replacement by a pronoun.

2.2. Debate on the role of predictability in pronoun usage

While the effects of the factors mentioned in the previous section are uncontested, the effect of referent predictability on pronoun production has been a topic of interest for many years and the results from previous studies have been mixed. Some studies find a difference in pronoun production between more predictable and less predictable referents (e.g., Arnold, 2001; Rosa and Arnold, 2017; Zerkle et al., 2017; Weatherford and Arnold, 2021), while others fail to find this difference (e.g., Kehler et al., 2008; Fukumura and Van Gompel, 2010; Mayol, 2018; Zhan et al., 2020). For an overview of the divergent conclusions drawn in previous work, see Table 6.1 in Chapter 6.

These varying outcomes have led to the proposal of models of pronoun production that posit different kinds of relationships between referent predictability and pronoun use. Two main models that embody different views on this matter are the Expectancy Hypothesis (Arnold, 1998, 2001, 2010) and the *strong* form of Bayesian Model (Kehler and Rohde, 2013), henceforth *Strong Bayes*.[1]

The Expectancy Hypothesis (Arnold, 1998, 2001; Arnold et al., 2007; Arnold, 2010; Arnold and Tanenhaus, 2011) posits that a listener's estimate of the likelihood that a particular referent will be mentioned next is closely associated with the activation level of that referent in the interlocutors' mental representation of discourse, that is, referent predictability is closely tied to referent accessibility. In traditional approaches to discourse anaphora, accessibility (often termed "salience", "prominence") denotes the activation level of a referent in the interlocutors' mental representation of discourse and is thought to play a critical role in determining speakers' choice of referring expressions (e.g., Givón, 1983; Gundel et al., 1993; Chafe, 1994; Brennan, 1995; Grosz et al., 1995). Specifically, highly accessible referents are more likely to be pronominalized, while less accessible references typically require more explicit expressions. Therefore, by positing an association between referent predictability and accessibility, the Expectancy Hypothesis predicts greater pronoun production for more predictable referents. Speakers are believed to calculate referent predictability as an estimate of accessibility, using more reduced forms, such as pronouns, for referents that are expected or highly predictable to their listener.

In contrast, Strong Bayes (Kehler and Rohde, 2013) posits that referent predictability, on the one hand, and pronoun production, on the other, are influenced by different contextual factors. Strong Bayes reflects an empirical observation that semantic and pragmatic factors influence predictability and accordingly affect the pronoun interpretation bias. However, the speaker's decision regarding the pronominalization of a referent is, according to Strong Bayes, insensitive to these factors. Instead, pronoun production is primarily influenced by grammati-

[1] There are two varieties of the Bayesian Model, a weak form and a strong form. The central claim of the Bayesian Model is embodied by its weak form, positing that an addressee interprets a pronoun by combining their estimate that the speaker is going to mention a particular referent next, with their estimate that the speaker would use a pronoun to mention this referent. We delve into this weak form in Chapter 4, where we examine both pronoun production and interpretation. However, this dissertation primarily focuses on the strong form, given its direct relevance to our principal research question and its direct contrast to the Expectancy Hypothesis.

cal factors or factors associated with information structure, such as subjecthood or topichood. Consequently, Strong Bayes predicts that a speaker's choice to pronominalize a referent will be unaffected by the set of semantic and pragmatic contextual factors that are known to influence referent predictability.

More details on formalizing the Expectancy Model (based on the Expectancy Hypothesis) and the Bayesian Model can be found in Section 4.1 of Chapter 4.

2.3. Referent predictability in passage continuation tasks

Previous empirical investigations concerning the effect of referent predictability on pronoun production have primarily employed passage continuation tasks with factorial designs and with semantic manipulations to vary predictability. In a typical 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 (5) below from Arnold (2001), and then provide a continuation based on how they expect the story to proceed, as for example in (6). In these studies, referent predictability is operationalized as the frequency of a referent being mentioned as the grammatical subject in the matrix clause of continuations. Thus, the referent that is most frequently re-mentioned first in continuations is considered to be the most predictable one.

(5) There was so much food for Thanksgiving, we didn't even eat half of it. Everyone got to take some food home. Lisa gave the leftover pie to Brendan. ______________

(6) Brendan loved pie and cakes and all manner of sweet things but didn't know how to bake.

To investigate the impact of referent predictability on pronoun use, researchers have utilized various manipulation methods to create scenarios involving contrasting levels of referent predictability. Among these methods, varying the main verbs and discourse connectives has been most widely adopted in the literature (e.g., Arnold, 2001; Fukumura and Van Gompel, 2010; Rohde and Kehler, 2014; Rosa and Arnold, 2017; Mayol, 2018; Zerkle and Arnold, 2019; Zhan et al., 2020; Hwang et al., 2022). Additionally, other less typical manipulation methods have also been used. For example, some studies manipulate the temporal structure of events (e.g., Ferretti et al., 2009), or the relative clause attached to direct objects (e.g., Kehler and Rohde, 2019).

Verb semantics. In examples like (5) above, verb semantics are often manipulated to vary referent predictability (e.g., Arnold, 2001; Fukumura and Van Gompel, 2010; Rohde and Kehler, 2014; Mayol, 2018; Zerkle and Arnold, 2019; Weatherford and Arnold, 2021). Specifically, the main verb "give" in the final prompt sentence "Lisa gave the leftover pie to Brendan" is a transfer-of-possession verb, which 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). The reason why these verbs have been useful to check the role of predictability in pronoun production is that they are divided into two subgroups with symmetric argument structures: Source-Goal verbs such as "give" in (7), and Goal-Source verbs like "catch" in (8).

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

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

Several passage continuation experiments have shown a consistent tendency for participants to more frequently refer back to the Goal referent than to the Source referent, in both types of verbs (e.g., Stevenson et al., 1994; Arnold, 2001; Rosa and 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 (7), participants tend to talk about what the Goal referent "Brendan" did next after receiving "leftover pie". To test how this expectation bias towards the Goal over the Source 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 (7) vs. "Lisa" in (8).

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 9-10), and they also elicit next-mention biases, but this time towards the Stimulus.[2]

[2] Some implicit causality verbs assign roles such as Agent and Patient (like "help"), or Agent and Evocator (like "criticize") (e.g., Goikoetxea et al., 2008). Such verbs can be broadly categorized into subject-biased and object-biased. In this thesis, I focus on Stimulus-Experiencer verbs to exemplify subject-biased implicit causality verbs and Experiencer-Stimulus for object-biased ones, due to their prominence in research and to maintain better comparability with transfer-of-possession verbs.

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

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

Studies have shown that these verbs induce biases towards the Stimulus antecedent 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 and Van Gompel, 2010; Ferstl et al., 2011; Rohde and Kehler, 2014; Mayol, 2018; Zhan et al., 2020). For example, in the case of “impress” in (9) there is a strong preference for referring back to the Stimulus, “David”, in the subject position, while for “admire” in (10), continuations also preferably refer to the Stimulus, “Linda”, which is in the object position. Therefore, “impress” would be biased towards the subject, while “admire” would be biased towards the object.

As in transfer-of-possession contexts, the bias towards the Stimulus over the Experiencer in implicit causality scenarios is used to test whether predictability affects pronoun production e.g., are there more pronouns produced referring to the Stimulus referent “David” in (9) than to the Experiencer referent “David” in (10)?

Discourse relations. Researchers have used discourse relations as another factor to manipulate referent predictability 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 and Van Gompel, 2010; Holler and Suckow, 2016; Hwang et al., 2022). For instance, while, as we have seen, speakers tend to continue Example (10) with the Stimulus, Linda, in an explanation (“because…”), they tend to continue Example (11) with the Experiencer, David, when talking about the result of the event (“so…”). The latter is the opposite of the default next-mention bias elicited by implicit causality verbs, which is towards the Stimulus as mentioned. Similar effects have been reported for transfer-of-possession verbs (Stevenson et al., 1994).

(11) 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 and Van Gompel, 2010; Holler and Suckow, 2016; Hwang et al., 2022; Hwang, 2023a), research of mine (Liao, 2022) and Hwang (2023b) extend the investigation of how discourse relations affect referent predictability to contexts beyond transfer-of-possession and implicit causality.[3] Specifically, Hwang (2023b) explores the role of connectives in facilitating a general sense of subject continuity and action continuity (see also Kehler 2002). Hwang (2023b) 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 (12) for a Korean example from Hwang (2023b). While Hwang (2023b) examined the production of Korean zero pronouns in discourse relations marked by distinct connectives, Chapter 3 and 4 of this thesis explore the production of overt English pronouns.

(12) Minswu-ka Hyenwu-wa palphyo cwunpi-lul
Minsu-NOM Hyunwoo-with presentation preparation-ACC
ha-ko/nuntey _______.
do-and/while _______
“Minsu prepared a presentation with Hyunwoo and (then)”/ “While Minsu was preparing a presentation with Hyunwoo”_______”

Other manipulations of referent predictability. While the manipulation of verb or connective semantics is a common strategy for varying the levels of referent predictability, alternative approaches also exist.

One such method, used by Kehler and Rohde (2019), involves manipulating relative clauses attached to direct objects in object-biased implicit causality contexts (see Example (13)). In their study, participants were found to produce fewer Explanation (e.g. those introduced by “because”) continuations for (13-a) than for (13-b) due to the relative clause in (13-a) already providing an explanation. As the object bias primarily arises within Explanation continuations, the decrease in the number of Explanation continuations for (13-a) thus yielded a difference in next-mention biases, with fewer object re-mentions for (13-a) relative to (13-b). In other words, “the patient” in (13-b) is more predictable than “the patient” in (13-a). The authors attribute this outcome to the fact that most coherence relations that participants could use in these contexts, other than Explanation, tended to exhibit a stronger subject bias (Kehler et al., 2008).

(13) a. The doctor reproached the patient **who never takes her medicine**. _______
b. The doctor reproached the patient **who came in at 3pm**. _______

In another study, Ferretti et al. (2009) explored the effects of manipulating the temporal structure of events (see Example (14)). They observed a higher propen-

[3] Liao (2022) is one of the studies included in this thesis and will be reported in Chapter 3.

sity for participants to re-mention the Goal rather than the Source in Source-Goal contexts across both perfective and imperfective conditions, (14-a) and (14-b). However, this bias towards the Goal was found to be reduced in (14-b). In other words, the Goal "Bob" is more predictable in (14-a) compared to (14-b). The authors attribute this finding to the higher salience of the Goal over the Source with respect to the end state of transfer-of-possession events (Stevenson et al., 1994; Arnold, 2001). The salience of the Goal is comparatively diminished in the imperfective condition, where the event is depicted as ongoing.

(14) a. John$_{source}$ **handed** a book to Bob$_{goal}$. ________
b. John$_{source}$ **was handing** a book to Bob$_{goal}$. ________

2.4. Referent predictability in upcoming referent guessing tasks

While most previous studies on referent predictability operationalize it as the frequency of a referent being mentioned next, other studies have approached this matter in other ways. Some studies employ information-theoretic measures such as surprisal and entropy in referent guessing tasks, where human subjects or computational models are tasked to guess which referent will be mentioned next (Tily and Piantadosi, 2009; Modi et al., 2017). In these studies, referent predictability is a function of the proportion of subjects that correctly guess the upcoming referent; or of a language model's certainty about the correct upcoming referent. Specifically, in a truncated corpus text, such as Example (15), adapted from Modi et al. (2016), participants are tasked with predicting the upcoming referent (indicated by XXXXXX). A lower percentage of guesses that align with the actual referred entity (e.g., the bathroom tub in the original corpus text (16)) corresponds to lower referent predictability. Similarly, a higher probability assigned by a language model to the upcoming referent co-referring with the actual referred entity, or greater certainty regarding this coreference, indicates higher referent predictability.

(15) I decided to take a bath yesterday afternoon after working out. Once I got back home, I walked to my bathroom and first quickly scrubbed the bathroom tub by turning on the water and rinsing it clean with a rag. After I finished, I plugged XXXXXX

(16) I decided to take a bath yesterday afternoon after working out. Once I got back home, I walked to my bathroom and first quickly scrubbed the bathroom tub by turning on the water and rinsing it clean with a rag. After I finished, I plugged **the tub** and began filling it with warm water

set at about 98 degrees.

Unlike passage continuation tasks, where participants choose both the referent to refer to and the form of reference to use in the continuation, referent guessing tasks examine the referential form used naturally in spoken or written language. The fundamental reasoning behind this method is that if predictability affects referential choice, referents that can be easily predicted based solely on the context are then more likely to be expressed with pronouns, while those that are harder to predict should be expressed with more explicit forms like names or descriptions. In this line of research, referent predictability is not manipulated in pairs of controlled contexts, and it is no longer a dichotomous variable (e.g., the more predictable referent "Bob" vs. the less predictable referent "Bob" in (14-a) and (14-b)). Instead, predictability is a continuous variable that is measured based on the likelihood of each referent in a sequence of naturally occurring passages being correctly guessed before a referring expression is revealed.

2.5. This thesis

Overall, in the existing body of research, the relationship between referent predictability and pronoun production has mostly been explored via passage continuation tasks, using carefully constructed materials featuring specific verb types as stimuli. This has resulted in a lack of representation of naturalistic language. Chapters 3 and 4 of this thesis work to bridge this gap by extending the empirical base to naturally occurring passages extracted from corpora, as well as to corpus passage continuations produced by human participants. On the other hand, Chapter 5 draws inspiration from a different approach seen in previous research–namely, the upcoming referent guessing tasks. Here, the focus is on exploring the potential of using a neural network model to perform the guessing task instead of humans, thus deriving reliable computational estimates of referent predictability. This alternative method enables a larger-scale exploration and analysis, without the constraints posed by the costs associated with human participants. In the meantime, as individual studies accumulate empirical evidence, the picture of the conclusions becomes increasingly mixed, which highlights the need for a more comprehensive examination of this relationship in empirical terms. Chapter 6 therefore presents a meta-analysis of the results of 20 independent studies on the topic to obtain a quantitative synthesis of the existing evidence. This approach enables us to estimate the overall effect size of referent predictability on pronoun production by pooling the results of different studies, and to examine the variability in effect sizes across studies.

Chapter 3

CORPUS ANALYSES

As mentioned in the previous chapter, prior empirical research has mostly employed passage continuation experiments to investigate whether predictability influences pronoun production (e.g., Arnold, 2001; Fukumura and Van Gompel, 2010; Rohde and Kehler, 2014; Zhan et al., 2020). These experiments employed targeted materials featuring specific verb types as stimuli. In this study, our objective is to investigate this question in more naturalistic contexts, using naturally occurring language. To this end, we automatically retrieve naturally produced texts from two corpora annotated in computational linguistic research.

We operationalize referent predictability in terms of referent re-mention frequency across certain specific 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 and Arnold, 2021). Under this premise, a referent (such as the grammatical subject) that is re-mentioned more frequently in a type of context compared to others is deemed more predictable for comprehenders within that type 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 Appendix A.

Therefore, our focus shifted toward investigating **discourse relations**. Recall from the previous chapter that 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 also 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 Appendix A, discourse relations can be expected to have a similar semantic general effect across contexts.

We identified three discourse relations for which we have specific hypotheses regarding their differing referent re-mention frequency, 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.[1] Examples illustrating these three relations can be found in Table 3.1.

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

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

All data processing and analysis code is publicly accessible through an OSF repository: `https://osf.io/n54sw/`.

3.1. Hypothesis and research question

We put forward the following hypothesis pertaining to the three relations of interest: Narration, Contrast, and Result.

Hypothesis: Predictability. Concerning the differing referent predictability 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 tendency 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 this effect to be particularly prominent in the case of Narration.

Research question: Relationship between predictability and pronoun production bias. We can explore our research question provided that the previous hypothesis is supported by the data. Regarding our question, recall that Strong Bayes predicts uniform pronoun production rates across the three relations —Narration, Contrast, and Result —despite the differing predictability of subject referents in their contexts. Alternatively, according to the Expectancy Hypothesis, discourse relations are predicted to influence not only referent predictability but also pronoun production, leading to a higher pronominalization rate for subject re-mentions in the Narration relation than in the other two relations.

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 however 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 sentences.

3.2. Corpora

3.2.1. 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 3.2.

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 3.2: English portion of OntoNotes: genres and corresponding sizes.

OntoNotes comes with rich manual annotations that enable its use for the purposes of the present study, exemplified in Table 3.3. 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 (17), 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.

(17) 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 3.3: Multiple layers of annotation in OntoNotes.

3.2.2. Rhetorical Structure Theory Discourse Treebank

Rhetorical Structure Theory Discourse Treebank (RST-DT; Carlson et al. 2001) 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 and Thompson, 1988).[2] 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 3.1, each leaf of the tree corresponds to an EDU. Adjacent EDUs are connected by discourse relations to form larger segments.

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

[2] Out of the 385 Wall Street Journal articles found in RST-DT, 277 are also included in the OntoNotes corpus. Unlike the analysis with OntoNotes, we use manual annotations to extract both implicit and explicit relations from RST-DT. This analysis thus presents evidence that is complementary to the previous one with OntoNotes.

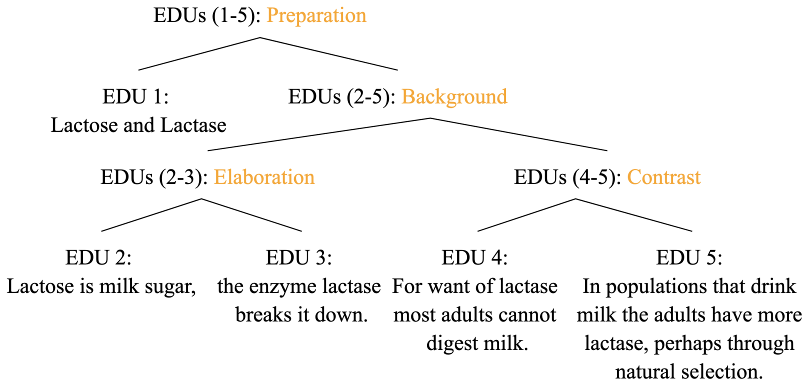

EDUs (1-5): Preparation
EDU 1:
Lactose and Lactase
EDUs (2-5): Background
EDUs (2-3): Elaboration
EDUs (4-5): Contrast
EDU 2:
Lactose is milk sugar,
EDU 3:
the enzyme lactase breaks it down.
EDU 4:
For want of lactase most adults cannot digest milk.
EDU 5:
In populations that drink milk the adults have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

Figure 3.1: Graphical representation of an RST analysis (own production using text from the RST website, www.sfu.ca/rst).

3.3. Method

3.3.1. Extraction of explicitly signaled relations from OntoNotes

To extract passages of discourse relations from the OntoNotes corpus, we relied on explicit connectives. The connectives used for our passage extraction are listed in Table 3.4. 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).[3] These selected connectives primarily signal the target relation rather than other relations, thus making them mostly non-ambiguous.[4]

We focused on cases of sentence-initial coordinating conjunction (inter-sentential), 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 con-

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

[4] The connective “so” often presents polysemy, which introduces some degree of ambiguity. However, we included it because it is very commonly used to indicate a 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 (‘IN’). Additionally, we consider the surrounding tokens within the extracted passages and exclude 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 3.4: Connectives used for passage extraction.

junctions 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.[5] 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 3.5 lists examples for the Result relation. Note that, unlike typical psycholinguistic experiments in this field, and like previous work using corpora (Arnold, 2001; Guan and 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.

3.3.2. 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

[5] 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 3.5: Automatic labeling paradigm of coreference types, exemplified with the Result relation.

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.

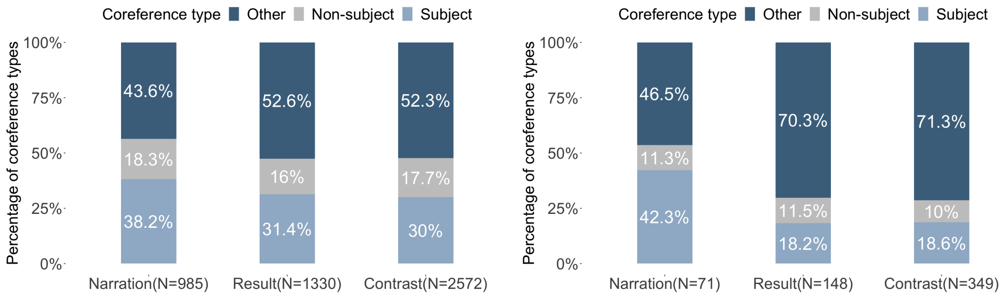
We selected RST relations that align approximately with the relations we extracted in OntoNotes, namely Narration, Contrast, and Result. The mapping between the original taxonomy of RST-DT and our categorization is provided in Table B.1, which can be found in Appendix B.1. For clarity, we maintain the 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*. For more details on our extraction strategy, see Appendix B.2.

3.4. Results

The results from both corpora support our hypothesis regarding predictability: the subject referent is more predictable in Narration than in other relations, as measured by the percentage of cases in which it is re-mentioned in the following clause. Figure 3.2 shows the distribution of coreference types by relation in OntoNotes and in RST-DT, where *subject* is higher in Narration than in Result and



(a) OntoNotes

(b) RST-DT

Figure 3.2: Coreference type by discourse relation in OntoNotes (left) and RST-DT (right).

Contrast (raw counts for each type are presented in Appendix B.3).

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.[6] In our analysis of the OntoNotes sample, random intercepts for verbs were included, along with random slopes for relation types by verb. 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.

The results of the statistical analysis are reported in Table 3.6 and 3.7 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 (thus more predictable) in the Narration relation compared to the Contrast relation. Pairwise comparisons show the subject referent is more frequently re-mentioned in Narration compared to Result (OntoNotes: β= 0.49, z = 3.07, p = 0.006; RST-DT: β= 1.29, z = 3.83, $p < 0.001$), while there is no difference between Contrast and Result (OntoNotes: β= 0.13, z = 0.99, p = 0.58; RST-DT: β= -0.01, z = -0.05, $p \approx 1$).

In terms of our research question, we find no evidence that predictability induced by discourse relations affects the choice of referring expression, as follows. Figure 3.3 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 de-

[6] We used the *glmer* function from the *lme4* package (v1.1-27; Bates et al. 2015) in R (R Core Team, 2019).

Effects		Estimate	SE	z	*p*
Intercept		-0.83	0.06	-13.86	
discourse relation	**Narration**	**0.36**	**0.11**	**3.13**	**0.002**
	Result	-0.13	0.13	-0.99	0.32

Table 3.6: Subject re-mention in OntoNotes: mixed-effects logistic regression model with the subject being re-mentioned as the dependent measure.

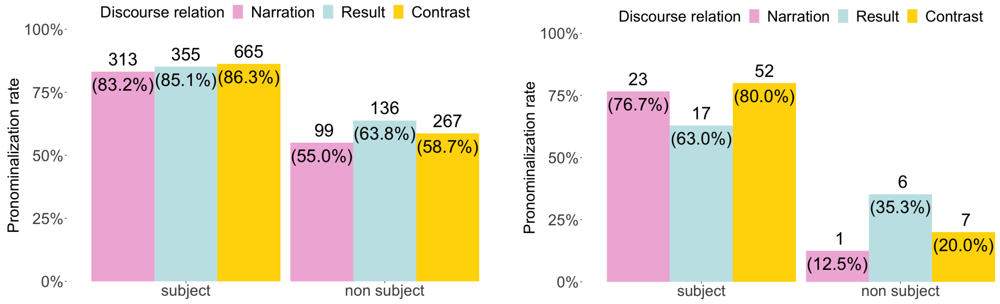
Effects		Estimate	SE	*z*	*p*
Intercept		-1.68	0.20	-8.45	
discourse relation	**Narration**	**1.31**	**0.30**	**4.33**	<**0.001**
	Result	0.01	0.26	0.05	0.96

Table 3.7: Subject re-mention in RST-DT: mixed-effects logistic regression models with the subject being re-mentioned as the dependent measure.

pendent 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. Random intercepts for the document ID were included to account for potential variations associated with e.g. author style. As shown in Table 3.8 and 3.9, 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 graphs (a) and (b) in Fig. 3.3) 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 Appendix B.4).

3.5. Discussion

Our corpus analyses reveal that while the predictability or next-mention frequency of subject antecedents varies across distinct discourse relations (Narration, Contrast, and Result), the rate of pronominalization remains consistent, in line



(a) OntoNotes

(b) RST-DT

Figure 3.3: Pronominalization rate of next mention by relation in OntoNotes (left) and RST-DT (right). 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		0.30	0.12	2.48	
discourse relation	Narration	-0.06	0.21	-0.31	0.76
	Result	0.20	0.20	1.02	0.31
coreference	**subject**	**1.66**	**0.17**	**10**	**<0.001**
Narration:subject		-0.003	0.28	-0.009	0.99
Contrast:subject		-0.31	0.28	-1.11	0.27

Table 3.8: 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 3.9: Pronominalization in RST-DT: Mixed-effects logistic regression model with the next mention being a pronoun as the dependent measure.

with the prediction of the 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 (the type of coherence relation between two sentences), while the likelihood of pronoun production appears to be insensitive to differences between discourse relations. Instead, pronoun production is primarily subject to the grammatical role: subject re-mentions were pronominalized significantly more often than non-subject re-mentions. This observation is consistent irrespective of whether relations are explicitly signaled by connectives or inferred by human annotators.

From a methodological perspective, our analyses showcase both the advantages and the challenges of using a corpus-based approach to investigate our linguistic question.

The clearest advantage of corpus-based approaches is the fact that they allow for the use of naturally produced language, and the examination of varied and diverse contexts. Most previous studies instead focused on a few verb types and used carefully controlled contexts to elicit continuations.

Another advantage of our methodology is the reduced cost associated with collecting human data and analyzing large datasets. By using existing resources like co-reference annotated corpora, we can automate much of the data extraction process, minimizing the need for manual annotation. This enables us to conduct larger-scale and more robust analyses than would be feasible with traditional experimental methods or earlier corpus work, which heavily relied on manual annotations (e.g. Arnold, 2001; Guan and Arnold, 2021). While programming challenges are involved in working with corpus data, these efforts can be readily recycled for related future research. For instance, we initially extracted passages featuring the two verb types. Later, when we shifted our focus to extracting passages related to discourse relations, we could do so without much additional effort, since we had already developed code for accessing different layers of annotations in the corpus.

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 lead to limited data availability in the corpus, making it difficult to conduct robust analyses. Future research could choose to focus on genres with linguistic structures more similar to the controlled contexts. For instance, if the study aimed to examine transfer-of-possession contexts, a corpus of sports commentary (e.g., football) might provide more relevant examples (e.g., expressions describing dynamic ball passing) than a general English corpus.

Another challenge is the fact that the stylistic features of a given corpus might not be tailored to the contexts of interest. For instance, we observe from a comparison between the two figures in Fig. 3.2 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 3.4(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 3.4(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.

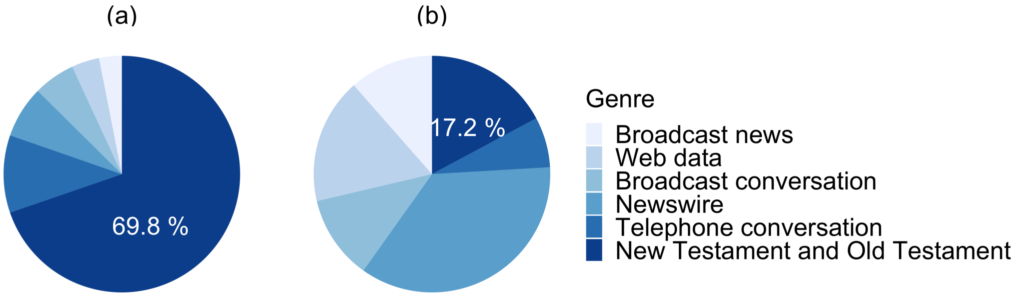


Figure 3.4: The left figure (a) presents the distribution of Narration coreference samples by genres in OntoNotes. The pie chart on the right (b) shows the genre distribution in OntoNotes.

We also note that relying solely on automated extraction methods in corpus analyses poses challenges due to the difficulty in controlling for all potential confounding factors. Unlike controlled experimental settings where researchers can manipulate and isolate specific factors, corpus data, being naturally occurring language, often requires careful and precise extraction using multiple layers of annotation to ensure accuracy. However, such annotations may be incomplete or entirely absent in corpora, as exemplified by the partial annotations of named entities in OntoNotes (see Footnote 1 in Appendix A.1.2), which can complicate the extraction process or lead to potential inaccuracies. On the other hand, controlling for more factors can result in a significantly reduced sample size or data scarcity, as demonstrated by our experience in extracting implicit causality and transfer-of-possession contexts (see Appendix A).

Furthermore, there is an ongoing challenge in corpus-based research related to the issue of data availability and corpus size. Although plenty of corpora are available for various languages and genres (particularly for English), there are fewer options that are both large and richly annotated with linguistic information. Researchers may face limitations when investigating less documented languages or very specific linguistic patterns. For instance, to examine our research question, a wealth of annotated data is required, including coreference chains, discourse relations, and syntactic parsing. However, even large-scale annotation projects like OntoNotes have limited resources for languages other than English. The annotated texts in Chinese and Arabic in OntoNotes, for example, are significantly smaller than their English counterparts. All this can limit the scope and applicability of corpus-based analyses.

To conclude, this corpus study contributes to the field in two ways. Methodologically, our work extends the empirical base to more naturalistic corpus passages and 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.

At a theoretical level, we show that discourse relations between clauses exhibit systematic patterns regarding predictability, which again broadens the empirical scope of the debate; and we find evidence consistent with the prediction of Strong Bayes that predictability induced by discourse relations does not influence pronoun production.

Building on our initial observation and exploration using ecologically valid corpus texts, we conducted the second study to obtain more robust experimental evidence supporting our findings regarding the predictability patterns and the relationship between predictability and pronoun production. In that study, we utilized a set of corpus passages as stimuli for a passage completion experiment with human participants. The following chapter reports this experimental study and offers a more in-depth discussion of the theoretical implications of our results.

Chapter 4

CORPUS PASSAGE COMPLETION EXPERIMENT

In our exploration of the relationship between predictability and pronoun production using corpus analyses, we found no evidence to suggest that discourse relations affect pronoun production, despite their effect on predictability.

As a relevant extension of our primary research question, we are also interested in examining how coherence-driven predictability affects pronoun interpretation. Specifically, we question whether referent predictability influences pronoun interpretation and production in a parallel manner. If there exists an underlying notion ("salience", "prominence", or "accessibility") that governs both pronoun production and interpretation, predictability would likely have a similar influence on both processes.

In this chapter, we conducted a controlled experiment, which allows us to examine both production and interpretation to compare three models of pronoun interpretation in addition to our primary research question. Specifically, we used a set of extracted corpus passages as stimuli in a story completion experiment with human participants. Following a commonly employed experimental paradigm in the field, we used both bare prompts and pronoun prompts. See examples (18) and (19).

(18) [Bare] John passed the comic to Bill. _______

(19) [Pronoun] John passed the comic to Bill. He _______

The bare-prompt data derived from the experiment provide the basis for examining the relationship between referent predictability and pronoun production biases. In addition, the same bare-prompt data also allow for generating predicted interpretation rates for the three models of pronoun interpretation in the literature –Bayesian, Mirror, and Expectancy models (see Section 4.1 for a more detailed discussion on these three models). We evaluate these models by comparing their predictions to the observations from the pronoun-prompt conditions.

All data processing and analysis code is publicly accessible through an OSF repository: `https://osf.io/n54sw/`.

4.1. Three models of pronoun interpretation

In addition to investigating the relationship between predictability and pronoun production, this chapter compares three models of pronoun interpretation: (1) the Mirror Model, which posits that listeners interpret pronouns as referring to the referents that the speaker chooses to mention with pronominal referring expressions; (2) the Expectancy Model, according to which listeners' interpretation bias toward a referent is their estimate of the probability that the referent will get mentioned next; and (3) the Bayesian Model, according to which pronoun interpretation is characterized as a combination of listeners' estimate that a referent will get mentioned next and listeners' expectation that the speaker will use a pronoun to refer to that referent.

We introduce these three models in more detail below.

4.1.1. Mirror Model

Common wisdom shared amongst discourse researchers indicates 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 4.1 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. 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

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

Table 4.1: Givenness Hierarchy
(Gundel et al., 1993)

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 (e.g., Rohde and Kehler, 2014; Bader and Portele, 2019; Zhan et al., 2020; Patterson et al., 2022), we adopt the term **Mirror Model** to refer to these types of approaches. In these approaches, pronoun production and interpretation align on the same notion of referent salience, essentially mirroring each other. We define this model as per Equation 4.1, as done in prior studies (e.g., Rohde and Kehler, 2014; Bader and Portele, 2019; Zhan et al., 2020; Patterson et al., 2022). The assignment operator "$\leftarrow$" is used to emphasize the fact that this model does not follow normative probability theory. In this equation, 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 $P(referent \mid pronoun)$ of 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, as in Eq. 4.2.

$$P(referent \mid pronoun) \leftarrow \frac{P(pronoun \mid referent)}{\sum\limits_{referent \in referents} P(pronoun \mid referent)} \tag{4.1}$$

$$P(referent \mid pronoun) \sim P(pronoun \mid referent) \tag{4.2}$$

4.1.2. Bayesian Model

Challenging the assumption of a unified salience notion in pronoun production and interpretation, an alternative proposal put forth by Kehler et al. (2008)

and Kehler and Rohde (2013) suggests a Bayesian framing for the relationship between pronoun interpretation and production. This model provides a plausible explanation for the asymmetries between pronoun production and interpretation observed in empirical studies (e.g., Rohde and 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 (20-a) towards either the subject "John" or the object "Bill" (a roughly 50/50 interpretation preference). However, for (20-b), 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".

(20) 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. 4.3. 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. Different next-mention biases will influence the interpretation of pronouns without directly affecting the production bias, giving rise to an asymmetry between the two.

$$P(referent \mid pronoun) = \frac{P(pronoun \mid referent)\, P(referent)}{\sum\limits_{referent \in referents} P(pronoun \mid referent)\, P(referent)} \tag{4.3}$$

More specifically, the pronoun interpretation bias, represented by $P(referent \mid pronoun)$, 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 expect Eq. (4.3) to hold approximately. This claim is known as the weak claim of the Bayesian Model, henceforth referred to as **Weak Bayes**.[1]

4.1.3. Expectancy Model

The third model is inspired by the Expectancy Hypothesis. As previously discussed in Section 2.2, the Expectancy Hypothesis suggests a direct link between listeners' estimate of the likelihood that a referent will be continued in the

[1] We have discussed the strong form of the Bayesian Model in Section 2.2.

discourse and the activation level of that referent in their mental representation (referent accessibility). As Arnold (2010) put it:

> *. . . discourse features influence accessibility by providing information about the predictability of upcoming reference. Accessible entities are those that are relatively likely to be mentioned in the current utterance – i.e., those with relatively high expectancy.*
>
> . . .
>
> *Under the communicative goal of referring, a plausible mechanism for expectancy is as a mechanism for discourse participants to coordinate accessibility. Expectancy describes how easily the comprehender will be able to retrieve the referent. Speakers could thus calculate expectancy as an estimate of accessibility to the listener.*

Therefore, the Expectancy Hypothesis predicts that pronoun interpretation is strongly influenced by the expectancy of a referent. Given this and following previous research (Rohde and Kehler, 2014; Bader and Portele, 2019; Zhan et al., 2020; Patterson et al., 2022), we define a third model of pronoun interpretation, termed as **Expectancy Model**. In this model, the pronoun interpretation bias primarily depends on the next-mention bias, as shown in Eq. 4.4, where the next-mention bias, denoted as $P(referent)$, is normalized by the probabilities of all possible referents. This implies that listeners tend to interpret a pronoun as referring to the referent that is most likely to be mentioned next by the speaker.

$$P(referent \mid pronoun) \leftarrow \frac{P(referent)}{\sum\limits_{referent \in referents} P(referent)} \tag{4.4}$$

4.2. Goals and hypotheses

Our investigation in this study is guided by two research questions.

One question addresses the central question of this dissertation: the relationship between referent predictability and pronoun production biases. Specifically, we examine whether pronoun production biases are affected by semantic and pragmatic factors—here, discourse relations—that are expected to influence next-mention biases. Concerning the next-mention biases, we formulate an identical hypothesis as the one presented in the earlier corpus analyses (see Section 3.1). That is, we expect a higher frequency of subject continuations in Narration than in the other two relations. Regarding our research question, the Strong Bayes predicted uniform pronoun production rates regardless of the differing subject's predictability across discourse relations, while the Expectancy Hypothesis suggests the opposite.

The second question pertains to evaluating which model for pronouns (i.e., Mirror, Bayesian, or Expectancy) best accounts for the interpretation rate. To achieve this, we compare the predictions of these three proposed models of pronoun interpretation against the observed interpretation biases.

4.3. Method

4.3.1. 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 the subject and object roles. The subsequent sentence began with an explicit connective, signaling either Narration, Contrast, or Result, as exemplified in (21).

(21) 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 (22-27). This design varied the Relation Type (Narration, Contrast, Result) and Prompt Type (bare vs. pronoun) within participants and passages. 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 the previous corpus study for passage extraction, see Table 3.4 in Chapter 3. 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.

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

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

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

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

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

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

In order to conceal the target manipulation, we incorporated 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 (28).

(28) 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.

4.3.2. Participants

200 individuals were recruited via the crowd-sourcing platform Prolific,[2] and participated in a 30-minute online experiment.[3] Participation was restricted to native English speakers residing in the United Kingdom. Prior to commencing

[2] URL: `https://www.prolific.co/`

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

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.

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.

4.3.3. Procedure

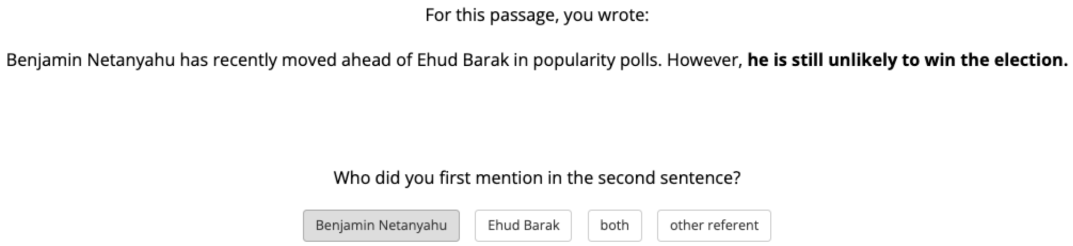
The study was presented using the JavaScript library *jsPsych* (De Leeuw 2015; version 7.1.2) and hosted on Pavlovia.[4] Participants were directed to the survey and randomly assigned to one of the six lists. Upon receiving written instructions, they proceeded to engage in a passage continuation task 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 4.1.

4.3.4. 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

[4] 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 4.1: Next-mention annotation in the format of four-choice questions.

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 et al., 2020), 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.*).[5] 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.

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.

To assess the quality and reliability of the automatic labeling process for referential form within bare-prompt completions, again a random sample of responses

[5] 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.

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 participants' annotations to be 93.4% and 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.

4.4. Analysis

The data were analyzed utilizing mixed-effect logistic regressions with centered predictors. We built three models: (1) a model of next-mention biases, which 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) a model of pronoun production biases, 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) a model of next mention with the fully crossed factors (Relation Type × Prompt Type) 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 simplified the random-effects structure until convergence was attained (Barr et al., 2013). We used likelihood ratio tests compare mixed-effects models that differed only in the presence or absence of a specific fixed effect or interaction.

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).[6] 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

[6] We started with the frequentist approach given the field's established familiarity with it.

were conducted to rule out any potential pathologies in the estimation process.[7]

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).

4.5. Results

Next-mention biases. Raw proportions of subject references by discourse relation and prompt type are shown in Figure 4.2. We first evaluated the next-mention biases (grey 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 4.2. Therefore, as observed in the previous corpus analyses (Chapter 3), the data once again supports Hypothesis 1: subject referents are more predictable in Narration than in Contrast and Result.

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 4.2: 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 Fig.4.3). 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

[7] 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.

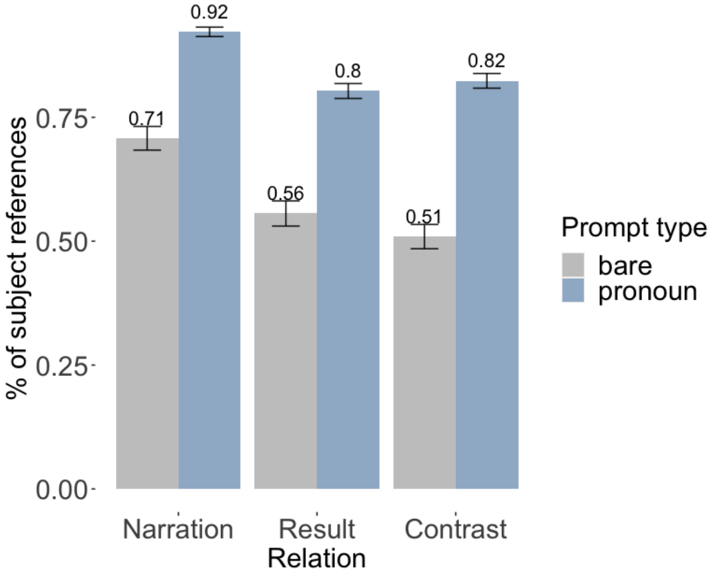


Figure 4.2: Proportion of subject references by discourse relations and prompt types. Error bars are standard errors over by-participant means.

subjects than non-subjects ($\beta = 2.71$, [2.11, 3.40]). As for Hypothesis 2, we found no evidence of an effect of Relation Type on pronominalization, in keeping with Strong Bayes as well as the findings from our previous corpus analyses. A model summary is presented in Table 4.3.

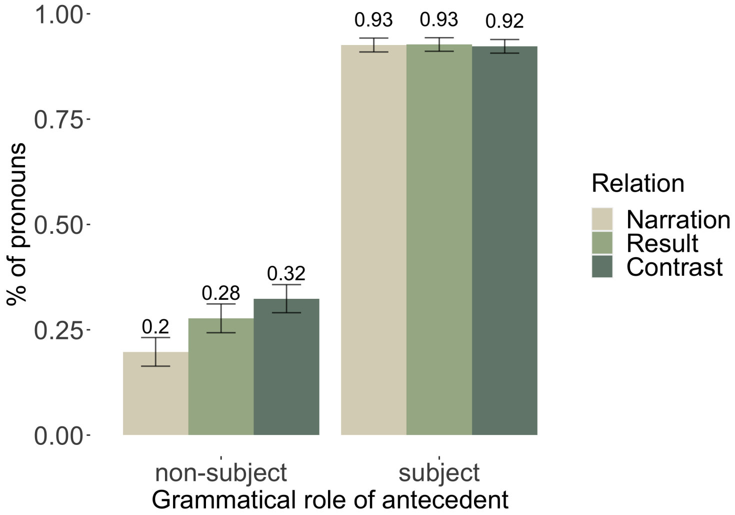


Figure 4.3: Pronominalization rates by grammatical roles and relation types.

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

Next mention with fully crossed factors. Next, we compare participants' interpretation biases measured in the pronoun-prompt condition (depicted in the blue bars of Figure 4.2) to their next-mention biases (already seen in the grey bars of Figure 4.2). 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 $\times$ Prompt Type. The summary of this model is presented in Table 4.4.

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. Post-hoc pairwise comparisons show that interpreting an ambiguous pronoun as the 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 4.4: Summary of logit mixed effect models of next mention with the fully crossed factors of Relation Type × Prompt Type (with all predictors centered).

4.6. Quantitative model comparisons

To address our research question on model evaluation, formulated in Section 4.2, 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 repeated in Table 4.5.

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

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

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

Table 4.5: Mathematical 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 smoothed stimulus-specific probability estimates as follows:[8]

$$\hat{P}(NP_j) = \frac{Count(NP_j) + W}{Count(NP1) + Count(NP2) + V \times W} \tag{4.5}$$

$$\hat{P}(\text{pronoun}|NP_j) = \frac{Count(NP_j \wedge \text{pronoun}) + 1}{Count(NP_j) + W} \tag{4.6}$$

Following previous studies (e.g., 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. We used two statistical metrics to assess how well the predictions of these models align with the observed interpretation biases: R-squared (R^2) and Mean Squared Error (MSE).

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 defined as shown in Eq. (4.7), where y_i, $\hat{y}_i$ and $\bar{y}$ represent, respectively, the observed interpretation preferences, the predicted interpretation preferences and the mean of interpretation preferences for the i-th stimulus. The variable n denotes the total number of stimuli. R^2 is therefore calculated by summing the squared residuals and squared differences between each observed interpretation rate and the mean. It is used to gauge the goodness of fit of these models, and the resulting 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 systematically underestimates or overestimates the observed biases, hence additional metrics are necessary to assess the models' performance.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.7}$$

We use Mean Squared Error (MSE), which calculates the average squared residual i.e., 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. When there is a systematic underestimation or overestimation in the predictions, the squared differences will increase the MSE. It is defined as (4.8), where y_i and $\hat{y}_i$ are respectively the observed and predicted interpretation preferences for

[8] In Eq. (4.6), "1" in the numerator represents the one pronominal form added for NP_j during the smoothing.

the i-th stimulus and n is the total number of stimuli. Hence, in the comparison of the three competing models of pronoun interpretation, the better the model, the larger the R-squared (R^2) and the smaller the Mean Squared Error (MSE) scores.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.8}$$

Results. The results of both metrics suggest that Strong Bayes outperforms the two competing models, as displayed in Table 4.6.[9] The Mirror Model demonstrates inferior performance when evaluated using the R^2 metric. However, when the evaluation is based on the MSE metric, the Expectancy Model exhibits comparatively weaker performance.

	Bayes	Expectancy	Mirror
R^2	**0.50**	0.43	0.03
MSE	**0.03**	0.10	0.06

Table 4.6: Results of statistical metrics for model comparisons. Best results boldfaced.

Figure 4.4 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.

4.7. Discussion

In this study, we conducted a passage completion experiment utilizing corpus passages as stimuli to investigate the relationship between predictability and pronoun production, as well as assessing three models of pronoun interpretation.

First of all, by manipulating discourse relation (Narration, Contrast, and Result) and prompt type (bare and pronoun), we obtained empirical evidence that

[9] 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.

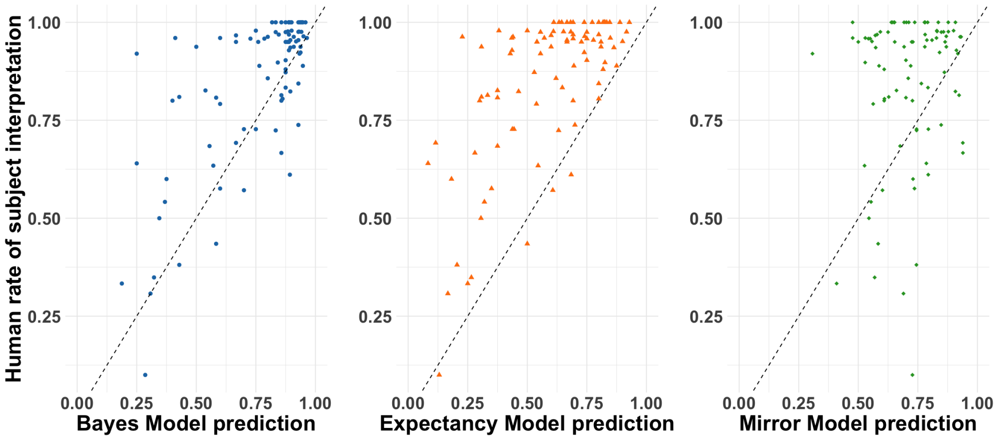


Figure 4.4: Model prediction vs. human rate for subject referents only. Each of the three plots shows the predicted pronoun interpretation rates from a different model. In total, there are 270 datapoints. Each datapoint represents a stimulus (30 items × 3 relations in the pronoun-prompt condition, 90 in total).

once again highlights a separation between the factors influencing next-mention bias and those influencing pronoun production bias, in line with the results from our previous corpus analyses and the prediction of Strong Bayes. Therefore, neither our corpus analyses nor passage continuation experiment yielded evidence supporting that predictability plays a role in pronoun production.

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 and Piantadosi, 2009). In fact, as mentioned in Chapter 1, 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 the continuation task and corpus analyses suggest 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 addressees 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 are the main drivers of a speaker's choice of referential form.

Regarding pronoun interpretation, our results favor the Bayesian Model over the two competing models. The central claim of the Bayesian Model, in its weaker form, is that the relationship between pronoun production and interpretation biases can be modeled 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. Specifically, they interpret a pronoun by combining their estimate that the speaker is going to mention a particular referent next, with their estimate that the speaker would use a pronoun to mention this referent, as formulated in Eq. 4.5 (a).

Interpretation biases, therefore, should reflect both next-mention biases and production biases. This was supported by our experimental findings. Concretely, the variation pattern in interpretation biases is the same as that in next-mention biases (i.e., larger interpretation and next-mention biases in Narration compared to Result and Contrast), 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 than the respective next-mention biases. Interpretation biases, therefore, should reflect both next-mention biases and production biases. This was supported by our experimental findings. Concretely, the variation pattern in interpretation biases is the same as that in next-mention biases (i.e., larger interpretation and next-mention biases in Narration compared to Result and Contrast), suggesting that the variation in next-mention biases shaped by the discourse relation manipulation became apparent through their influence on interpretation biases. Moreover, when comparing across relations, interpretation biases are much more skewed towards the previous subject than next-mention biases. This is likely driven by the expected influence from the pronoun production biases favoring the grammatical subject.

Further, we analyzed the predictions of the Bayesian Model along with two other models—the Expectancy Model and the Mirror Model. The Bayesian Model 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 4.4, 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.

Note that the alignment between Bayesian Model predictions and observed interpretation rates is not perfect either. It tends to underestimate the interpretation bias towards the previous subject, albeit less so than the Expectancy Model. For 16 out of 90 stimuli (10 Narration, 3 Result, and 3 Contrast), the observed interpretation rate maxes out at 1, indicating that all participants interpreted the pronoun referring to the previous subject. However, with our smoothing method, the predicted interpretation rates by all three models are impossible to reach 1. Notably, the Bayesian Model's predictions cluster in the upper right corner, indicating better accuracy in these extreme instances.

To conclude, both our experimental study and previous corpus analyses show that discourse relations between clauses exhibit systematic patterns regarding predictability. In both studies, we provide evidence for the prediction of Strong Bayes that pronoun production is unaffected by predictability induced by discourse relations. Moreover, model-comparison analyses of the experimental data provide robust evidence that the Bayesian model overall made more accurate predictions for pronoun interpretation than production and next-mention biases separately. Hence, our data provide broad support for the Bayesian Model of pronoun production and interpretation.

Chapter 5

COMPUTATIONAL MODELING

In line with the majority of psycholinguistic investigations concerning the relationship between predictability and pronoun production, both our corpus analyses and passage completion experiment have relied on next-mention frequency as an approximation for referent predictability. Neither approach, however, scales well due to the constraints posed by the limited availability of annotated data and the costs of involving human participants.

Following the long tradition of using computational models trained on large amounts of data as proxies for different aspects of human cognition, in this study we extend to the referential level previous research that uses computational models to obtain predictability scores (Hale, 2001; Smith and Levy, 2013). Specifically, we enable a computational model to predict a discourse entity without seeing its linguistics realization, in a similar manner as humans do in the upcoming referent guessing tasks (Tily and Piantadosi, 2009; Modi et al., 2017). The predictions generated by our computational model then serve as the basis for estimating predictability. If successful, this strategy would allow investigations into this question in a wider set of contexts than in psycholinguistic experiments and corpus analyses. Given the novelty of this methodology within the field, the next section of this chapter will first enhance the background context and provide an overview of related work.

We make the code used to carry out this study available at `https://github.com/amore-upf/masked-coreference`.

5.1. Background

5.1.1. Large language models built upon deep neural networks

Probabilistic language models are computational tools that assign conditional probabilities to linguistic elements, such as words, phrases, or syntactic structures. As illustrated in Figure 5.1, they can be employed to produce a probability distribution over the vocabulary, which captures the likelihood of each word or token occurring in a given linguistic context. In recent years, Natural Language Processing (NLP) has witnessed remarkable advancements, largely attributed to the development of Large Language Models (LLMs), which leverage deep neural networks (see below) to acquire remarkable linguistic skills.

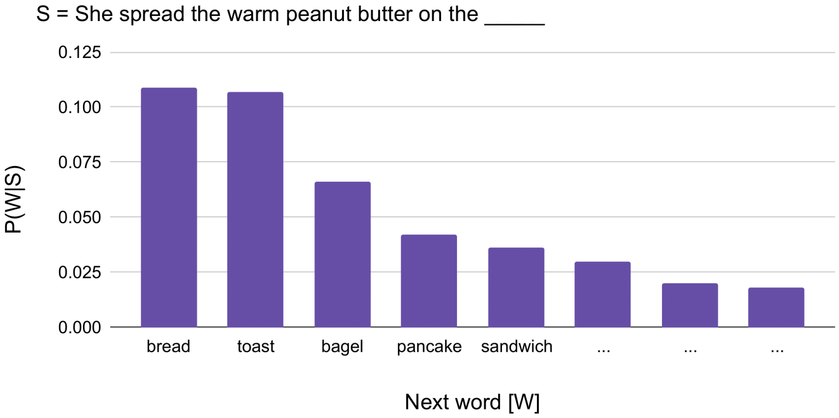


Figure 5.1: Probability distribution showing the likelihood of each word following the sequence “She spread the warm peanut butter on the”.

Modern artificial neural networks are inspired by the structure and function of biological neurons in the human brain. They have played a crucial role in the current Artificial Intelligence revolution. These networks consist of interconnected artificial neurons (or units), which are organized into layers to form a complex system capable of learning and representing data. The use of deep neural networks, characterized by multiple hidden layers between the input and output layers (as illustrated in Figure 5.2), has become increasingly popular for tackling complex tasks such as NLP. As Boleda (2020) aptly put it, “Neural networks are a type of machine learning algorithm, recently revamped as deep learning (LeCun et al., 2015), that induce representations of the data they are fed in the process of learning to perform a task”.

At the time the experiments in this chapter were carried out, one of the most revolutionary and influential LLMs was BERT (short for Bidirectional Encoder

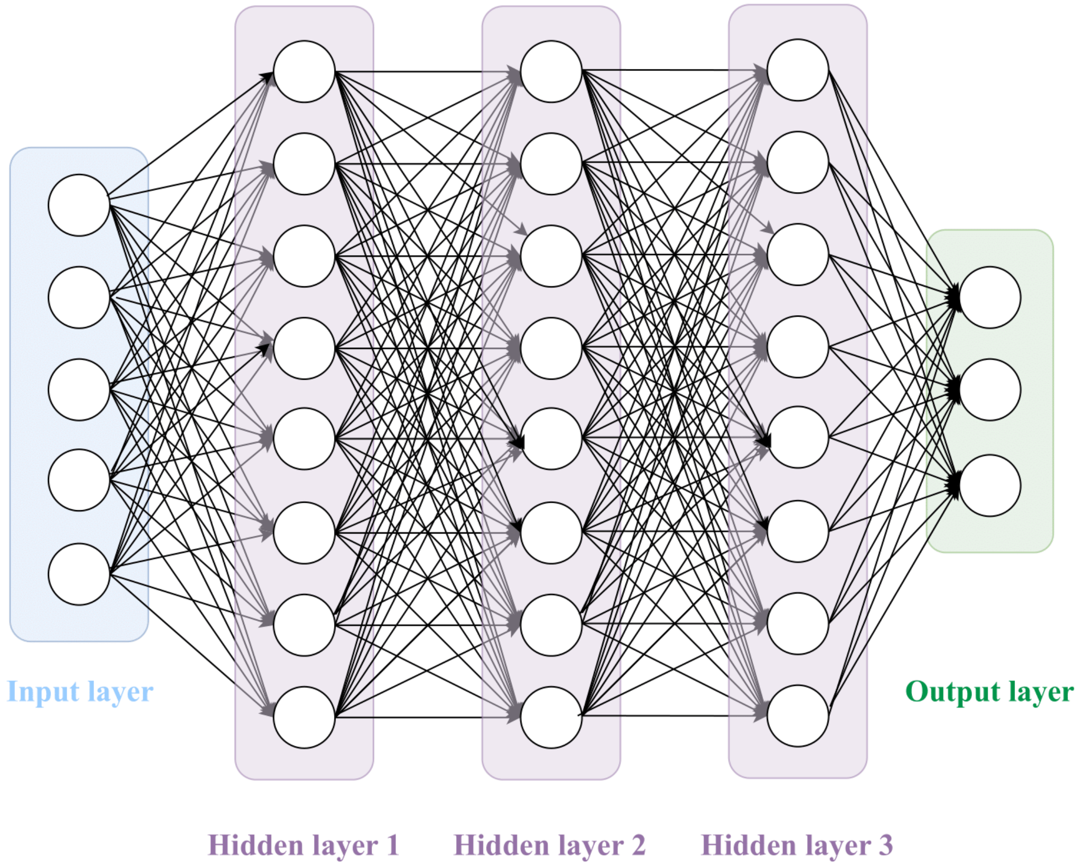


Figure 5.2: An illustration of a deep neural network, comprising multiple layers of interconnected nodes (representing artificial neurons), with an input layer, three hidden layers, and an output layer.

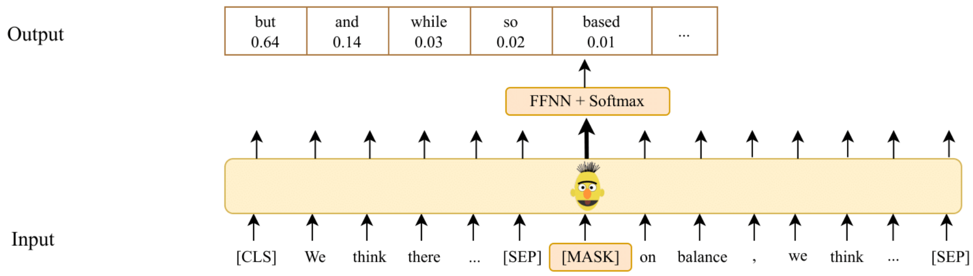


Figure 5.3: Illustration of the mask-filling task with BERT given bidirectional contexts. [CLS] and [SEP] are sentence boundary tokens commonly used in BERT pre-training for the Next Sentence Prediction task.

Representations from Transformers; Devlin et al. 2019). BERT is pre-trained on vast amounts of unlabeled web data, using two primary tasks: masked language modeling and next-sentence prediction. Masked language modeling involves predicting missing tokens, i.e., [MASK] tokens, given bidirectional context, as shown in Figure 5.3. This helps BERT learn contextual word representations. Next-sentence prediction, which involves determining whether two sentences are in sequence, enhances BERT's comprehension of the relationships between sentences and longer-term dependencies across sentences. Additionally, BERT uses attention mechanisms to determine how much attention each word in a sequence should pay to every other word in the same sequence, allowing it to capture contextual relationships between words. Specifically, BERT learns multi-head attention, where multiple attention mechanisms operate in parallel, enabling it to capture a broader range of relationships between words.

One advantage of LLMs like BERT is that they can be fine-tuned on specific downstream tasks, enabling them to adapt to new tasks without requiring extensive retraining. This approach is rooted in the idea that a well-trained general model can serve as a generic model of language. Researchers can build upon this foundation by fine-tuning the pre-trained model for a particular task. In analogy, an LLM can be likened to a master chef with vast culinary knowledge, while fine-tuning represents the process of perfecting a specific recipe based on their existing expertise, without needing to start learning to cook from scratch.

5.1.2. Computational estimates of predictability

Research into expectation-based human language comprehension has widely adopted these models to generate predictions about upcoming words given the words seen so far in a context (Armeni et al., 2017). Predictability scores obtained from language models have been shown to correlate with measures of cognitive cost at the lexical and the syntactic levels (Smith and Levy, 2013; Frank et al., 2013). In these studies, predictability is typically measured with solely the preceding context, as shown in Figure 5.1 (e.g., Levy, 2008). However, more recent work has also looked at predictability calculated in a bidirectional setup (i.e., based on both the previous and following contexts), like we do in this study. For example, Pimentel et al. (2020) used surprisal (i.e. negative log-probability; also see Section 5.4) calculated from a language model which takes both left and right pieces of context (e.g., She spread the warm peanut butter on the _____ in the kitchen). They studied the trade-off between clarity and cost and reported a tendency for ambiguous words to appear in highly informative contexts.

Other work also used computational estimates of referent predictability. In Orita et al. (2015), they were used to explain speakers' choice of referential form. Their measures of predictability, however, were based on simple features like frequency and recency. In a related vein, Modi et al. (2017) built computational models aimed at predicting the upcoming referent. Their models combined shallow linguistic features (e.g., recency, frequency, grammatical function) and script knowledge (common-sense knowledge about everyday event sequences). This approach allowed them to disentangle the role of linguistic and common-sense knowledge, respectively, on the human referent predictions gathered in their study. More recent research assessed the ability of autoregressive language models[1] to mimick referential expectation biases that humans have shown in the psycholinguistic literature (e.g., Upadhye et al., 2020). Davis and van Schijndel (2021) extended this assessment to non-autoregressive models, like the ones we use in this study, and reported results consistent with prior work in autoregressive models, showing that these models exhibited biases in line with existing evidence on human behavior, at least for English.

5.1.3. Automatic coreference resolution

In our study, we propose an adaption of a coreference resolution system in order to estimate referent predictability. The goal of a standard coreference resolution system is to group mentions in a text according to the real-world entity they refer to (Pradhan et al., 2012). As shown in Example (29), coreference resolution systems are trained to find mentions that refer to the same entity in the text, resulting in clusters of mentions such as <My sister, She> and <a dog, him>.

(29) My sister has a dog. She loves him a lot.

[1] Autoregressive models generate sequences (like words) one step at a time, and each step is conditioned on the previous steps. In comparison, non-autoregressive models generate sequences without relying on previously generated parts of the sequence. They aim to predict all parts of the sequence simultaneously.

Several deep learning approaches to coreference resolution have been proposed in the field of NLP, such as cluster-ranking (Clark and Manning, 2016) or span-ranking architectures (Lee et al., 2017). We focus on span-ranking models, which output, for each mention, a probability distribution over its potential antecedents, as shown in Figure 5.4.

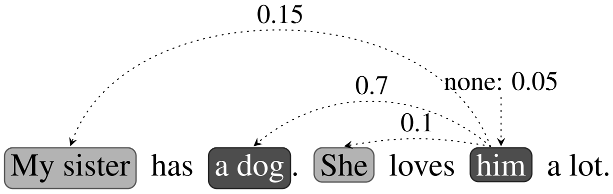
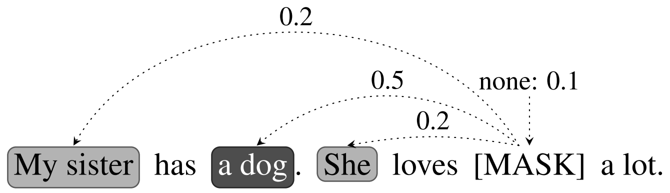


Figure 5.4: An example of probabilities from span-ranking coreference models showing the likelihood of "him" coreferring with each potential antecedent. "none" indicates "him" as a new entity's first mention.

We rely on an existing state-of-the-art coreference resolution system based on the SpanBERT language model (Joshi et al., 2020), henceforth SpanBERT-coref. It builds directly on the coreference systems of Joshi et al. (2019) and Lee et al. (2018), the main innovation being its reliance SpanBERT in place of the previous language models (respectively) BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018). We give more details about the system in Section 5.2.

Both BERT and its extension which underlies SpanBERT-coref, SpanBERT, are pretrained on masked language modeling (Devlin et al., 2019): a percentage of tokens is *masked* – substituted with a special token `[MASK]`; the model has to predict these tokens based on their surrounding context. This strategy encourages the models to develop a fuller understanding of the context and relationships between words. Unlike BERT, which masks individual tokens, for training SpanBERT, random contiguous sets of tokens –*spans*– are masked, and additionally, the model is trained to predict the entire masked spans based on tokens at the start and end of the boundary of the span (known as Span Boundary Objective). This training methodology resulted in SpanBERT performing better than BERT when used for adapting to tasks involving spans of tokens, like coreference resolution.

Standard coreference models, however, cannot be directly used to model predictability, because they are trained with access to both the context of a referring expression and the expression itself, as in Figure 5.4. Instead, we aim at obtaining predictability scores that are only conditioned on the context, following the definition of predictability used in psycholinguistics. To this end, we minimally modify a state-of-the-art coreference system (Joshi et al., 2020) to also carry out what we call *masked coreference resolution* (Figure 5.5): computing referent probabilities without observing the target mention. We show that the resulting model retains standard coreference resolution performance, while yielding a better estimate of human-derived referent predictability than previous attempts (see Section 5.3).



$P(E_{\text{[MASK]}} = \{\text{My sister, She}\}) = P(\text{antecedent}_{\text{[MASK]}} = \text{My sister}) + P(\text{antecedent}_{\text{[MASK]}} = \text{She}) = 0.4$
$P(E_{\text{[MASK]}} = \{\text{a dog}\}) = P(\text{antecedent}_{\text{[MASK]}} = \text{a dog}) = 0.5$
$P(E_{\text{[MASK]}} = \text{new}) = P(\text{antecedent}_{\text{[MASK]}} = \text{none}) = 0.1$

Figure 5.5: An example of deriving referent probabilities from masked coreference resolution predictions. “E” represents “entity”.

5.2. Methods

5.2.1. The SpanBERT-coref system

We use the best coreference resolution system SpanBERT-coref from Joshi et al. (2020), which is based on the language model SpanBERT-base (12 hidden layers of 768 units, with 12 attention heads). We make no changes to its architecture, only to the data on which it is trained, by masking some mentions.

In SpanBERT-coref, each span of text is represented by a fixed-length vector computed from SpanBERT token representations, obtained considering a surrounding context window[2] of maximum 384 tokens.[3] From span representations, a mention score is computed for each span (s_m: how likely it is to be a mention), and a compatibility score is computed for each pair of spans (s_a: how likely it is that they corefer). These scores are aggregated into a score s for each pair of spans (Eq. 5.1). For each mention, a probability distribution over its candidate antecedents (previous mentions and ‘none’) is then derived, determining coreference links (Eq. 5.2). The complete system is trained (and the underlying language model finetuned) on the English portion of the coreference-annotated OntoNotes

[2] A surrounding context window refers to a specific range or span of text that the system considers when computing token representations.

[3] A span representation is the concatenation of its start token representation, end token representation, and a weighted sum (attention) over all of its token representations.

5.0 dataset (the same corpus used in Chapter 3; Weischedel et al. 2013), which we also use to train the model on the *masked coreference resolution* task in this chapter.

$$s(x, y) = s_m(x) + s_m(y) + s_a(x, y) \tag{5.1}$$

$$P(\text{antecedent}_x = y) = \frac{e^{s(x,y)}}{\sum_{i \in \text{candidate}_x} e^{s(x,i)}} \tag{5.2}$$

5.2.2. Training on masked coreference

We model entity predictability in terms of a probability distribution over entities given a masked mention. The probability that a mention x refers to the entity e – $P(E_x = e)$ – can be computed as the sum of the antecedent probabilities of the mentions of e in the previous discourse (M_e):

$$P(E_x = e) = \sum_{i \in M_e} P(\text{antecedent}_x = i) \tag{5.3}$$

However, in SpanBERT-coref this probability is conditioned both on the mention and its context (the model has observed both to compute a prediction), whereas we need a distribution that is only conditioned on the context.

To achieve this, we introduce *masked coreference resolution*, the task of determining the antecedent of a mention that has been replaced by an uninformative token `[MASK]`. The task, inspired on masked language modeling (Devlin et al., 2019), is illustrated in Figure 5.5. Note that SpanBERT-coref can be directly used for masked coreference, since its vocabulary already includes the `[MASK]` token. However, since the system was not trained in this setup, its predictions are not optimal, as we show in Section 5.3. Therefore, we train a new instance of SpanBERT-coref adapted to our purposes – $\mathbf{M}_\text{m}$. To train $\mathbf{M}_\text{m}$, we mask a random sample of mentions in each document by replacing them with a single `[MASK]` token.[4] The percentage of mentions masked is a hyperparameter (we test 5%-40%). Note that, this way, the model is optimized to identify the correct antecedents of both masked and unmasked mentions, such that it retains standard coreference capabilities (see Section 5.3).

5.2.3. Evaluation

In this section, we describe how we obtain and evaluate predictions on masked and unmasked mentions, respectively. While our primary focus is on the pre-

[4] Masked spans are re-sampled in a document each time this document passes through the algorithm for the system training. [MASK] replaces the entire span of each mention, independently of its length. We verified that the use of one [MASK] token did not bias $\mathbf{M}_\text{m}$ to expect single-token mentions such as pronouns; see Figure C.5 in Appendix.

dictions made for masked mentions—since we use them to derive estimates of referent predictability—we also examine the quality of predictions for unmasked mentions to offer a comparative perspective.

Unmasked mentions. When discussing predictions for unmasked mentions, we are talking about predictions made by the system while it has access to both the mention itself and its surrounding context. These predictions are obtained through the standard coreference resolution task, as in Figure 5.4. To do this, we simply feed the documents in our dataset to the system as they are, without masking any mentions. The system then links each mention to the cluster of its most probable antecedent or creates a new cluster if it determines that there is no suitable antecedent, thus producing the predicted clusters in the process.

To evaluate the performance on this standard coreference resolution, following the CoNLL-2012 shared task (Pradhan et al., 2012), we report the averages of the MUC, B^3 and CEAF metrics in precision, recall and F1, respectively. All these metrics focus on the quality of the predicted clusters of mentions (i.e., coreference chains) compared to the gold ones (human-annotated clusters). MUC (Vilain et al., 1995) considers a cluster of mentions as linked mentions, wherein each mention is linked to at most two other mentions (a mention is linked to its antecedent, if one exists, and it may also be linked to its subsequent re-mention, if applicable). It counts the number of link modifications required to make the clusters predicted by the system identical to the gold clusters. B^3 (Bagga and Baldwin, 1998), on the other hand, calculates precision (the proportion of correct elements in the predicted cluster) and recall (the ratio of correct elements in the predicted chain to the total number of elements in the gold cluster) for every mention. These values are then averaged over all mentions. The CEAF metric (Luo, 2005), instead, assumes that each gold cluster should only be mapped to one predicted cluster, and vice versa. It evaluates the similarity between predicted clusters and gold clusters.

Masked mentions. As previously mentioned, the system can be employed to perform the standard coreference resolution task on the documents in their original form, which allows us to obtain predictions for all unmasked mentions simultaneously. However, when it comes to obtaining predictions on masked mentions, this approach is not feasible. This is due to the fact that masking all mentions in a document conceals all potential antecedents within the document, making the coreference resolution task impossible.

To obtain predictions on masked mentions, our approach needs to follow the setup illustrated in Figure 5.5. In this setup, we mask only the target mention which we intend to predict the referent of, while keeping the surrounding context, including potential antecedents, accessible to the system. This implies the need to mask different mentions within a document individually and pass the same document with varying mentions masked as separate inputs to the system. We apply this method specifically when we compare the model's predictions with those made by humans from the Modi et al. (2017) dataset, as elaborated in Section 5.3.3. For larger-scale analyses on the OntoNotes test data, including system evaluations and analyses of referential form, we employ an intermediate strategy. Specifically, for each document, a partition of the mentions, where each subset is *maskable* if, for each mention, none of its antecedents nor surrounding tokens (50 on either side) are masked. We generate one version of each document for each subset of masked mentions. We compute predictions for each document version separately, and collect antecedent assignments for the masked mentions, thereby obtaining masked predictions for each mention in the document.[5] This approach seeks to strike a balance between computational efficiency (we want to avoid masking only one mention per document at a time) and potential interference among the masked mentions.

Given that only a portion of mentions within a document are masked, the conventional metrics (MUC, B^3 and CEAF) employed for standard coreference resolution evaluation are not suitable. This is because these metrics evaluate predictions at the cluster level, conflating performances on masked and unmasked mentions. For masked mentions, our primary objective is to evaluate the quality of antecedent probabilities for a given target mention. These probabilities serve as proxies for referent predictability. We do not focus on the quality of the cluster the target mention is linked to, which also depends on the model's prediction for other mentions. Therefore, for masked mentions, we primarily evaluate the system in terms of *antecedent prediction*. We measure antecedent precision, recall and F1, where a model's prediction for a mention is correct if it assigns the largest probability to a true antecedent (an antecedent belonging to the correct mention cluster), or to "none" if it is the first mention of an entity. We use F1 on antecedent prediction as the criterion for model selection during development. In addition, we also evaluate antecedent prediction for unmasked mentions for comparison.

5.2.4. Using gold mention boundaries

As mentioned earlier in Section 5.2.1, SpanBERT-coref is trained to accomplish two tasks jointly: detecting mentions (i.e., their span boundaries) and identifying coreference links between them.

We are interested only in the latter task, such that this dual-task setup poses challenges to our purpose. This is because errors can arise in the latter task due to

[5] For hyperparameter tuning on development data, we use a faster but more coarse-grained method, described in Appendix C.1.

undetected antecedents or missed predictions for certain mentions because these mentions are not detected in the former task.

To address the challenge of mention detection and reduce the impact of noise, we propose an approach for setting up the system. In this approach, during the deployment of the system on test data, the system directly utilizes the known gold mentions to identify coreference links and does not need to predict mention boundaries in the first place. Importantly, this strategy does not necessitate any modifications to the model training process.

More specifically, we provide the system with the annotated gold mention boundaries and ensure that these boundaries are the only candidate spans considered for the system's output predictions. We then nullify the influence of mention scores on antecedent predictions. This is accomplished by setting mention scores to zero for all mentions (i.e., $s(i, j) = s_a(i, j)$; see Equation 5.1). By doing this, any potential error made by the model regarding mention detection will not affect the antecedent prediction.

To assess the system's performance, we compare its behavior when using predicted mention boundaries versus gold mention boundaries. Our expectation is that the latter setup will result in better proxies of referent predictability.

5.2.5. Context

The SpanBERT-coref system uses both preceding and following contexts of mentions when performing coreference resolution. Concretely, for a target mention, the system considers only mentions from the preceding context as candidate antecedents. However, the mention representation used to compute predictions takes into account both the preceding and following context of the mention. This approach aligns with the bidirectional nature of SpanBERT, the underlying language model of SpanBERT-coref, and also aligns with the current state of the art in coreference resolution.

In our experiments, we maintain this aspect of the model without modification. This sets our approach apart from the upcoming referent guessing tasks of Tily and Piantadosi (2009) and Modi et al. (2017), where participants were provided only with the preceding context of a mention. While most studies on referent predictability typically focus on the preceding context of a mention, it is worth considering that addressees can also take the following context into account when interpreting referring expressions (van Deemter, 1990; Song and Kaiser, 2020). The speaker may take this into account for their choice of mention form. Thus, the notion of referent predictability we model is to be understood not in the sense of anticipation in time, but in informational terms (in line with Levy and Jaeger 2007): how much information the context provides vs. how much information the mention needs to provide for its intended referent to be understood.

Our decision to incorporate the following context is not only theoretically motivated but also practical. Most state-of-the-art coreference resolution models are built on bidirectional architectures, and retaining this aspect enables us to obtain better estimations of referent predictability with minimal changes to these systems. While we could have chosen a more controlled approach by feeding the model different inputs based on the mention of interest (e.g., only the preceding context up to each mention), as opposed to a fixed-size window, this approach would be computationally inefficient, especially during training. This is because predictions related to mentions within a document cannot be obtained by simply passing the document to the system once; rather, it necessitates feeding distinct versions of the document, each cropped at the end boundaries of the specific mention of interest. Moreover, there's no guarantee that the model would adapt well to this setup, considering it wasn't originally trained in this manner.

In our experiments with Modi et al. (2017)'s data, as detailed in Section 5.3.3, we compute predictions based on the previous context only. However, in practice, we have observed that this setup often yields predictions that are less human-like compared to the bidirectional setup, despite being more aligned with how the data were collected.[6] This can likely be attributed to the mismatch between the model's training and deployment conditions. We leave the exploration of different kinds of context to future work.

5.3. Evaluation

5.3.1. General results

Table 5.1 reports the results of evaluation on OntoNotes test data for both $\mathbf{M}_{\mathrm{u}}$ (the standard SpanBERT-coref coreference system) and our variant $\mathbf{M}_{\mathrm{m}}$, trained with 15% of mentions masked in each document.[7] The table reports results for both model-predicted and gold mention boundaries; the latter are always higher, as expected. For unmasked mentions, we provide results both for standard coreference resolution and per-mention antecedent predictions (ANTECEDENT); for masked mentions, only the latter is applicable (see Section 5.2.3).

On unmasked mentions, the two models perform basically the same. This means that masking 15% of mentions during training, which was done only for $\mathbf{M}_{\mathrm{m}}$, does not interfere with the ability of the system on ordinary, unmasked coreference resolution. On masked mentions, both models perform worse, which is

[6] Experiments on antecedent prediction on OntoNotes led to similar results.

[7] Models trained masking between 10%-35% of mentions achieved comparable antecedent accuracy on masked mentions on development data. Among the best setups, we select for analysis the best model in terms of coreference resolution.

		UNMASKED MENTIONS						MASKED MENTIONS		
		COREFERENCE			ANTECEDENT			ANTECEDENT		
	boundaries	P	R	F1	P	R	F1	P	R	F1
M$_u$	predicted	.78	.77	.77	.86	.82	.84	.42	.39	.4
	gold	.91	.85	.88		.90			.50	
M$_m$	predicted	.78	.76	.77	.86	.82	.84	.69	.69	**.69**
	gold	.91	.86	.88		.91			**.74**	

Table 5.1: Results on OntoNotes test data (English): document-level coreference resolution (only with unmasked mentions; CoNLL scores) and antecedent prediction (both unmasked and masked mentions); P, R, F1 = precision, recall, F1 scores (when using gold mention boundaries on antecedent prediction, P = R = F1). BUC, M^3 and CEAF scores are reported in Appendix C.2.

expected because it is a more difficult task: Without lexical information about the mention, the models have less information to base their prediction on. Still, both models provide correct predictions in a non-trivial portion of cases. **M**$_u$, which did not observe any masked mentions during training, achieves .5 F1 for gold mentions. A random baseline gets only .08, and selecting always the immediately previous mention or always "no antecedent" obtain .23 and .26, respectively. Thus, it seems that **M**$_u$ retains some ability to compute meaningful representations for masked tokens from pretraining, despite not seeing `[MASK]` during training for coreference resolution. Nevertheless, in line with our expectations, training on masked coreference is beneficial: **M**$_m$ improves substantially over the results of **M**$_u$, with .74 F1. This means that even without lexical information from the mention itself, 74% of referents are correctly identified, i.e., predictable on the basis of context alone.

5.3.2. Results by mention type

Figure 5.6 breaks down the antecedent precision scores of **M**$_m$ by mention type. From now on we look only at the setup with gold mention boundaries, though the trends are the same for predicted mentions (reported in Appendix C.2). We distinguish between proper names (e.g., "Kamala Harris"), full noun phrases (NPs; e.g., "the tall tree") and pronouns (e.g., "she", "that"). For completeness, in Appendix C.2 we report the results considering a more fine-grained distinction.

The figure shows that for predicting the antecedent of masked mentions, pronouns are the easiest (.81), followed by proper names (.71), and full NPs (.66) are the hardest. Put differently, pronouns are used in places where the referent is

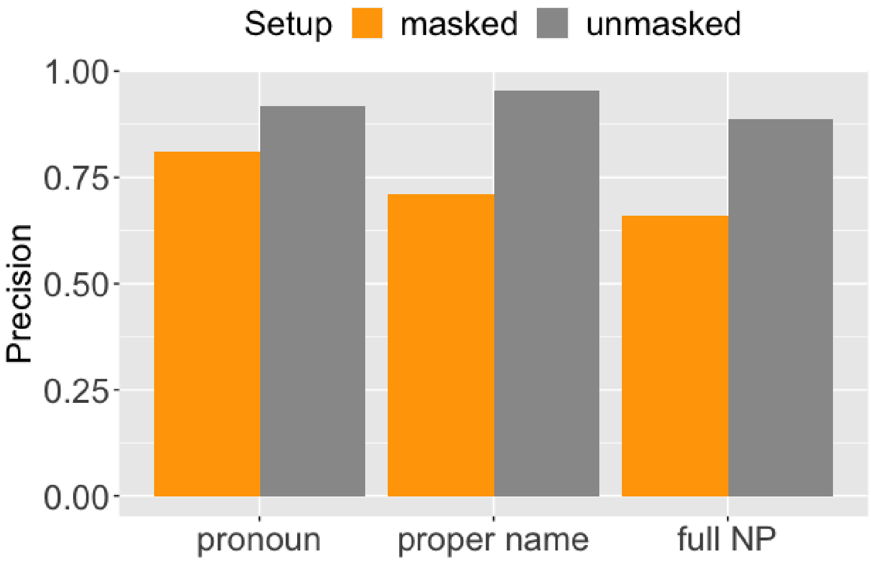


Figure 5.6: Antecedent precision for $\mathbf{M}_\mathrm{m}$ across different mention types, for masked and unmasked mentions.

the most predictable, full NPs when the referent is the least predictable. Table 5.2 shows examples of predictions on masked mentions with different mention types.

For unmasked mentions, instead, proper names are the easiest (.96; names are typically very informative of the intended referent), and full NPs (.89) are only slightly more difficult than pronouns (.92). Hence, the pattern we see for masked mentions cannot be a mere side-effect of pronouns being easier to resolve in general (also when unmasked), which does not seem to be the case. Instead, it provides initial evidence for the expected relation between referent predictability and mention choice, which we will investigate more in the next section.

5.3.3. Comparison to human predictions

We assess how human-like our model $\mathbf{M}_\mathrm{m}$ behaves by comparing its outputs to human guesses in the cloze-task data from Modi et al. (2017). Subjects were asked to guess the antecedent of a masked mention in narrative stories while seeing only the left context (182 stories, $\sim$3K mentions with 20 guesses each; see Ex. (15) for an example). To evaluate the model's estimates, we follow Modi et al.'s approach, and compute Jensen-Shannon divergence to measure the dissimilarity between a model's output and the human distribution over guessed referents. The lower the divergence, the better. $\mathbf{M}_\mathrm{m}$ achieves a divergence of .46, better than Modi et al.'s best model (.50), indicating that our system better approximates human expectations. Appendix C.2 provides further results and details.

5.4. Predictability and mention form

context	mention
(1) Judy Miller is protecting another source [...] Let me get a response from Lucy Dalglish. I think it's very obvious from what Judy wrote today **[MASK]** is protecting somebody else. ✓	she
(2) This child [...] felt particularly lonely and especially wanted his father to come back. He said that he was sick one time. **[MASK]** worked in Guyuan ×	his father
(3) One high-profile provision [...] was the proposal by Chairman Lloyd Bentsen of the Senate Finance Committee to expand the deduction for individual retirement accounts. **[MASK]** said he hopes the Senate will consider that measure soon ✓	Mr. Bentsen
(4) Sharon Osbourne, Ozzy's long-time manager, wife and best friend, announced to the world that she'd been diagnosed with colon cancer. Every fiber of Ozzy was shaken. **[MASK]** had to be sedated for a while. ×	he

Table 5.2: Examples of correct and incorrect predictions by $\mathbf{M}_{\text{m}}$ (with gold mention boundaries) on masked mentions; model's prediction underlined, correct antecedent with dotted line.

The previous section assessed the effect of our masked training method on model quality. We believe that the model predictions are of high quality enough that we can use them to test the main hypothesis regarding the relation between predictability and mention choice. Following previous work (see Section 5.1), we define predictability in terms of the information-theoretic notion of **surprisal**: the more predictable an event, the lower our surprisal when it actually occurs. Given a masked mention x with its true referent e_{true}, surprisal is computed from the model's output probability distribution over entities E_x (Eq. 5.3), given the context c_x:

$$\text{surprisal}(x) := -\log_2 P(E_x = e_{\text{true}} \mid c_x)$$

Surprisal ranges from 0 (if the probability assigned to the correct entity equals 1) to infinity (if this probability equals 0). Surprisal depends only on the probability assigned to the correct entity, regardless of the level of uncertainty between the competitors. As Tily and Piantadosi (2009) note, uncertainty between competitors is expected to be relevant for mention choice, e.g., a pronoun may be safely used if no competitors are likely, but risks being ambiguous if a second entity is somewhat likely. Tily and Piantadosi (2009) and, following them, Modi et al. (2017) took this uncertainty into account in terms of entropy, i.e., *expected* surprisal. We report our analyses using entropy in Appendix C.3, for reasons of space and because they support essentially the same conclusions as the analyses using just surprisal.

We check whether surprisal predicts mention type (Section 5.4.1) and whether it predicts mention length (number of tokens; Section 5.4.2). All analyses in this section use the probabilities computed by $\mathbf{M}_{\mathrm{m}}$ with gold mention boundaries.

5.4.1. Surprisal as a predictor of mention type

For this analysis, in line with previous studies, we consider only third person pronouns, proper names and full NPs with an antecedent (i.e., not the first mention of an entity). For the OntoNotes test data this amounts to 9758 datapoints (4281 pronouns, 2213 proper names and 3264 full NPs). Figure 5.7 visualizes surprisal of masked mentions grouped by type, showing that despite much within-type variation, full NPs tend to have higher surprisal (be less predictable) than pronouns and proper names.

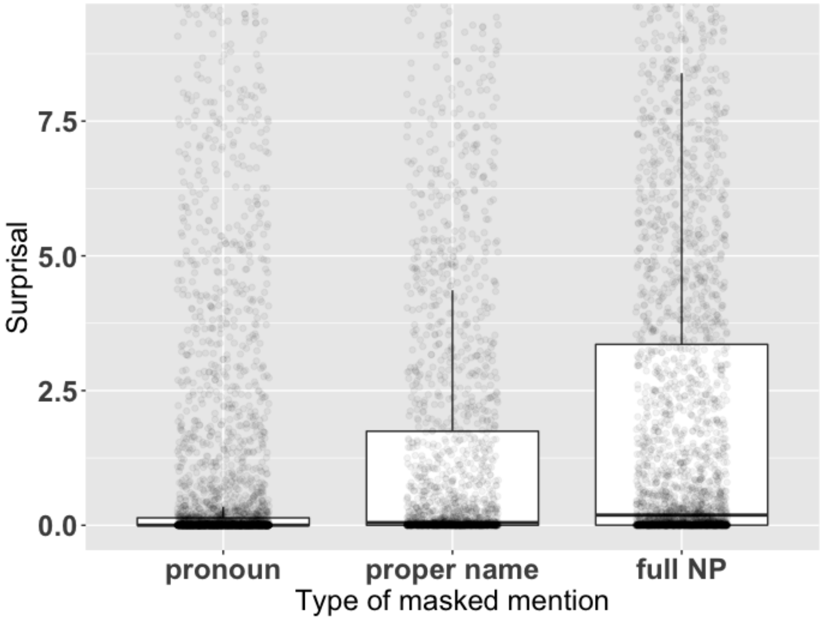


Figure 5.7: Surprisal and mention type. The limits on the y axis were scaled to the 95th percentile of the data to visualize the variability better.

To quantify the effect of predictability on mention type, we use multinomial logistic regression, using as the dependent variable the three-way referential choice with pronoun as the base level, and surprisal as independent variable.[8] The results of this surprisal-only regression are given in the top left segment of Table 5.3. The coefficients show that greater surprisal is associated with a higher probability assigned to proper names ($\beta = .31$) and even more so full NPs ($\beta = .47$); hence pronouns are used for more predictable referents. Since surprisal was standardized, we can interpret the coefficients (from logits to probabilities): e.g., adding one standard deviation from mean surprisal increases the predicted probability of a proper name from .23 to .25, and of a full NP from .33 to .42, decreasing the probability of a pronoun from .43 to .34.

Next, following Tily and Piantadosi (2009) and Modi et al. (2017), we test whether predictability has any effect over and above shallower linguistic features from the literature that have been hypothesized to affect mention choice. We fit a new regression model including the following features as independent variables alongside surprisal:[9] **distance** (number of sentences between target mention and its closest antecedent); **frequency** (number of mentions of the target mention's referent so far); closest **antecedent is previous subject** (i.e., of the previous clause); **target mention is subject**; closest **antecedent type** (pronoun, proper name, or full NP). The results are shown in the bottom left segment of Table 5.3.[10] We verified that the incorporation of each predictor improved goodness-of-fit, using the Likelihood Ratio (LR) chi-squared test (with standard .05 alpha level; full tables with LR Chi-squared test are reported in Appendix C.3.). Surprisal improved goodness-of-fit ($p_{\chi^2} << 0.001$): it contributes relevant information not captured by the shallow features alone. At the same time, however, now surprisal is not anymore predictive of the distinction between pronouns and proper names, as found by Tily and Piantadosi (2009) – only of the distinction between pronouns and full NPs (see significance values of the predictor "surprisal" for the two left columns of Table 5.3).

If we conceive of the shallow features as possible confounds, our results show that predictability still affects mention choice when controlling for these. Alternatively, we can take the shallow features to themselves capture aspects of predictability (e.g., grammatical subjects tend to be used for topical referents, which are therefore expected to be mentioned again), in which case the results show that

[8] We use the *multinom* procedure from the library *nnet* (Venables and Ripley, 2002). Continuous predictors were standardised, thus allowing for comparison of coefficients.

[9] The result of this simultaneous regression as regards the predictor surprisal will be identical to what the result would be of a hierarchical regression where surprisal is the last added predictor (Wurm and Fisicaro, 2014).

[10] We visualize the comparison of observed to predicted types using a ternary plot, see Figure C.6 in Appendix C.3.

these features do not capture all aspects.

As for the shallow features themselves, we find that pronouns are favoured over proper names and full NPs when the referent has been mentioned recently, in line with the idea that the use of pronouns is sensitive to the local salience of a referent. Moreover, pronominalization is more likely if the previous mention of the referent was itself a pronoun. There is also a strong tendency to reuse proper names, perhaps due to stylistic features of the texts in OntoNotes: in news texts, proper names are often repeatedly used, plausibly to avoid confusion, as news articles often introduce many entities in a short span; in the Bible, the use of repeated proper names is especially common for the protagonists (e.g. Jesus). Lastly, we find the well-known *subject bias* for pronouns: pronouns are more likely than full NPs or proper names when the referent's previous mention occurred in subject position.

Overall, the results corroborate the finding in Tily and Piantadosi (2009) that full NPs are favoured, and pronouns and proper names disfavored, when surprisal is higher; and extend their finding, based on newspaper texts only, to a larger amount of data and more diverse genres of text (news, magazine articles, weblogs, religious texts, broadcast and telephone conversation).

		Predicting mention type								Predicting mention length			
		Proper name				Full NP							
		β	s.e.	z	p	β	s.e.	z	p	β	s.e.	t	p
Intercept		-.63	.03	-23.8	-	-.26	.02	-10.9	-	1.87	.02	80.8	-
surprisal		.31	.03	9.6	*	.47	.03	16.4	*	.25	.02	10.7	*
Intercept		-.24	.07	-3.6	-	.04	.07	.6	-	1.81	.05	40.1	-
distance		3.13	.12	25.4	*	3.10	.12	25.2	*	.17	.02	7.1	*
frequency		.09	.03	3.1	*	-.13	.03	-3.8	*	-.13	.02	-5.4	*
antecedent	subject	-1.31	.09	-13.9	*	-1.10	.08	-13.7	*	-.51	.06	-8.5	*
mention	subject	.07	.07	1.0	.3	-0.50	.06	-7.7	*	.04	.05	.8	.4
antecedent type	name	1.78	.08	22.8	*	.41	.09	4.6	*	-.21	.06	-3.2	*
	NP	-.17	.08	-2.2	*	1.18	.06	18.1	*	.42	.06	7.5	*
surprisal		.05	.04	1.5	.1	.23	.03	7.8	*	.17	.02	7.4	*

Table 5.3: (left) Two Multinomial logit models predicting mention type (baseline level is "pronoun"), (right) two linear regression models predicting mention length (number of tokens) of the masked mention, based on 1) surprisal alone and 2) shallow linguistic features + surprisal. * marks predictors that are significant at the .05 alpha level.

5.4.2. Surprisal as a predictor of mention length

If pronouns are favoured for more predictable referents due to a trade-off between information content and cost, one would expect to find similar patterns using graded measures of utterance cost, instead of flattening it to coarse-grained distinctions across mention types. In this subsection we use the number of tokens as such a measure (Orita et al., 2015). The average number of tokens per mention in our dataset is (of course) 1 for pronouns, 1.67 for proper names and 3.16 for full NPs.

We fit linear regression models with *mention length* in number of tokens as the dependent variable (or number of characters, in Appendix C.3), and, again, surprisal with and without shallow linguistic features as independent variables. The right segment of Table 5.3 presents the results, indeed showing an effect of mention length. In the surprisal-only model, moving up by one standard deviation increases the predicted mention length by .25 tokens (or 1.40 characters, see Table C.5 in Appendix C.3). Grammatical function and type of the antecedent are still strong predictors, with surprisal again making a contribution on top of that: mentions that refer to a more surprising referent tend to have more words. Figure 5.8 visualizes this trend between surprisal and predicted mention length.

Single-token pronouns dominate the lower end of the output range, raising the question of whether predictability still makes a difference if we exclude them, i.e., fit regression models only on the non-pronominal mentions. Our results support an affirmative answer (see Table C.4 and C.6 in Appendix C.3): the more surprising a referent, the longer the proper name or full NP tends to be.

5.5. Discussion

In this study, we investigated the relationship between referent predictability and the choice of referring expression using computational estimates of the former. To derive these, we adapted an existing coreference resolution system to operate in a setup resembling those of upcoming referent guessing tasks employed in psycholinguistics. Using computational estimates of semantic expectations allowed us to scale and expedite analyses on a large dataset, spanning different genres and domains.

Contrary to the findings in the previous two studies, our results in this study point to a trade-off between clarity and cost, whereby shorter and possibly more ambiguous expressions are used in contexts where the referent is more predictable. We found this both when grouping mentions by their morphosyntactic type as well as when considering their length. Referent predictability seems to play a partially overlapping but complementary role on referential choice with features affecting the salience of an entity, such as its recency, frequency, or whether it was last mentioned as a subject. This points to the open question as to whether salience or accessibility can actually be reduced to predictability (Arnold, 2001; Zarcone et al., 2016).

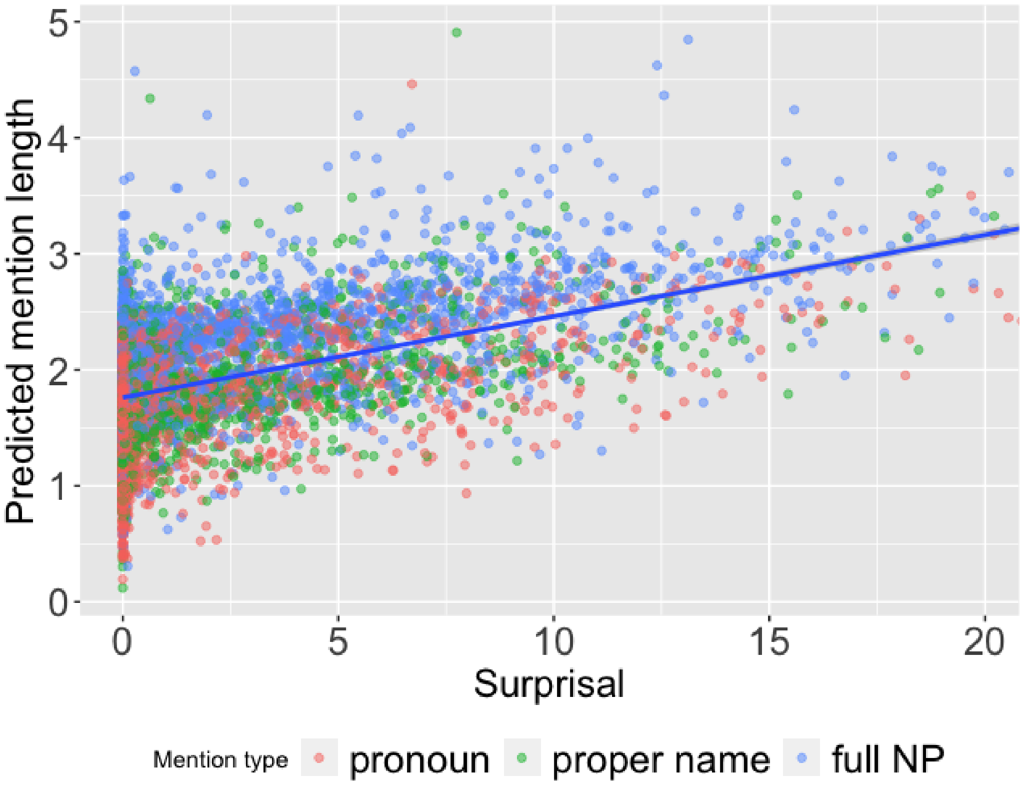


Figure 5.8: Trend between surprisal and predicted mention length by the linear regression model, visualized by adding a smoothing line comparing only the outcome with the variable surprisal.

Our bidirectional setup (using both the preceding and following contexts) is not directly comparable to that of some of the related work in terms of the amount and scope of context given for prediction. Referents are predicted with only the preceding context in previous work, both in psycholinguistic and computational approaches, while our model gives predictions based on both the preceding and following contexts. A key hypothesis shared between previous studies and our own is that speakers tend to avoid redundancies between the informativeness of context and that of referring expression. Given this, our results highlight a significant question that needs further investigation. What type of context influences referential choice? Is it only the preceding context or does the following one also play a role? And to what extent does either side influence this choice? Leventhal (1973) raised a similar question concerning word intelligibility in sentences and found that participants delayed the decision about a test word presented in noise until the entire sentence was encoded, and that the context after the target word was more facilitating to its intelligibility. Song and Kaiser (2020) also showed that comprenhenders actively utilized post-pronominal information in pronoun resolution. The use of a computational model provides flexibility to compare predictions using different amounts of context, and could shed light on how the previous and following context affect mention choice. Future work could also use unidirectional models, which allow for a setup more like the one adopted by prior work for ease of comparison, if requirements on the quality of performance can be met.

We hope that our work will foster the use of computational models in the study of referential choice and linguistic questions in general. This approach offers several advantages. Firstly, by using computational models, we can significantly reduce the need for recruiting subjects for experiments, saving costs associated with human participants. In addition, once a model is trained for a specific task, it can be easily applied to new, unlabeled data. This allows us to explore phenomena beyond those considered here without much additional effort. While our work focuses on English, the same methodology can be applied to more languages besides English (provided the availability of coreference resources; for instance, Arabic and Chinese are included in OntoNotes 5.0).

Nonetheless, obtaining high-quality linguistic data to train models in the first place can be a challenge. The coreference resources such as the dataset we used in this chapter are not available for many languages. There are also other challenges associated with using computational models, including the difficulty in interpreting model predictions due to the "black box" nature of neural networks. While these model can approximate human predictions to some extent, there's no guarantee that the models are truly behaving like humans. This poses an obstacle to deriving meaningful linguistic insights.

Relevant future venues are more fine-grained classifications of NPs (such as indefinite vs. definite), the effect of referent predictability on processing (McDonald and MacWhinney, 1995), and the kinds of context examined in psycholinguistic experiments (e.g., different discourse relations, verbs with contrasting referent biases; Rohde and Kehler 2014; Mayol 2018).

Chapter 6

META-ANALYSIS

The current body of literature, as well as the previous three studies, presents a complex and inconclusive picture, which highlights the need for a more comprehensive and robust examination of this relationship. In this last study, we address this issue by conducting a meta-analysis. Specifically, we use statistical methods to synthesize the results of 20 independent studies on the topic to obtain a quantitative synthesis of the existing evidence and to estimate the overall effect size of referent predictability on pronoun production. The studies comprise 26 experiments, of which 8 found that predictability affects pronoun production and 14 did not find this effect.

One notable advantage of meta-analytic methods is the major increase in statistical power achieved by pooling data from multiple studies. This is particularly valuable when the literature is mixed or when studies have small sample sizes and low statistical power (e.g., Borenstein et al., 2009; Cumming, 2012). In the present meta-analysis, data from over 1,145 unique participants were included. This starkly contrasts with any individual experiment on the topic, with none exceeding 100 participants. Using a meta-analysis, we speak to our research question from a more robust vantage point, enabled through a larger body of empirical evidence than that of any study taken individually.

Moreover, meta-analyses enable us to examine the variability in effect sizes across studies. This can help identify potential moderators (i.e., variables that influence the strength or direction of the relationship between referent predictability and pronoun production) that may explain the heterogeneity in effect sizes across studies, such as task differences or characteristics of the stimuli.

In fact, meta-analysis is a statistical technique commonly employed in disciplines that rely heavily on quantitative research, such as medicine, psychology, economics, biology, and environmental science. In contrast, its application in linguistics remains uncommon, primarily due to the field's smaller size, historical reliance on non-experimental methodologies, and the traditionally restricted availability of experimental data. In psycholinguistics, the application of meta-analyses has also been relatively limited. Notwithstanding, there has been an increase in recent years. For example, Mahowald et al. (2016) conduct a meta-analysis to estimate the effect of syntactic priming in production; Jäger et al. (2017) do so for the effect of retrieval interference in sentence comprehension.

All data processing and analysis code developed for this meta-analysis is available at: `https://osf.io/qyahc/`.

6.1. Scope of this study

While a range of factors can influence referent predictability (see Chapter 2), this meta-analysis focuses on predictability that is primarily driven by semantic and pragmatic factors that play a role in establishing the coherence of the discourse. That is, we investigate whether pronoun production is sensitive to semantically driven contextual biases that have been shown to influence expectations about the upcoming referent. This question is the central question of this thesis and lies at the heart of the difference between the Expectancy Hypothesis and the strong form of the Bayesian Model.

In contrast, the impact of grammatical factors and information structural factors, such as topichood, on the likelihood of re-mention and pronoun production is more firmly established (see, for instance, Centering Theory; Brennan et al. 1987; Brennan 1995). Empirical research within this domain mostly corroborates the widely recognized influence of grammatical role on pronoun production, with more pronouns produced referring to subjects than non-subjects.

Therefore, the studies that are relevant to our meta-analysis are those that manipulate specific semantic and pragmatic factors. We have excluded studies that exclusively manipulated focushood in their experimental designs (Kaiser, 2010) or employed less typical manipulations that remain less well-defined in the literature, such as the information status and uniqueness status of referents (Brocher et al., 2018), order of mention (Fukumura and van Gompel, 2015), frequency of referent nouns (Lau and Hwang, 2016), indefiniteness by case-marking in Turkish (Özge et al., 2016), referent animacy (Fukumura and Van Gompel, 2011), referent specificity by pe-marking in Romanian (Chiriacescu and von Heusinger, 2010), and social status of referents (Vogels, 2019).

Lastly, our meta-analysis also excludes studies that operationalized predictability using information theoretic notions such as surprisal or entropy (e.g., our study in Chapter 5). The main reason for excluding this line of work is that these studies require a different measure of the effect size of predictability than the majority of other studies, which use next-mention frequency as a way of quantifying referent predictability (see Section 6.2.3 for the calculation of effect size). As discussed in Section 2.4, predictability is a dichotomous variable in story continuation tasks but a continuous one in referent guessing tasks. Excluding studies that use information-theoretic measures also limits the potential for methodological differences across studies to confound our findings.

6.2. Method

6.2.1. Study selection criteria

To present as comprehensive a picture of current research as possible, we conducted a literature search using a combination of keyword and forward methods (see Harari et al. 2020 for a summary of study identification methods). The full process is visualized in Figure 6.1.

As a first step (*Search* in Fig. 6.1), in January 2023, we located relevant studies in the academic search engine Google Scholar,[1] using the following two features: (i) they contain a combination of keywords: *refer* AND *pronoun* AND (*completion* OR *production*); (ii) they cite at least one of the four representative articles on this topic: Arnold (1998), Arnold (2001), Kehler et al. (2008), and Kehler and Rohde (2013). This resulted in 776 articles for which we next conducted an abstract screening (*Screening* in Fig. 6.1).

6.2.2. Inclusion criteria

We established the following criteria to include studies in our meta-analysis. First (criterion 1), the study uses a typical manipulation type in the field, with a focus on coherence-driven predictability (see Section 6.1 for study scope). Thus, the data collected in the study enables comparison of referring expression usage for more predictable referents and less predictable referents, while controlling for their grammatical role. Second (criterion 2), the candidate study codes both choice of next-mention and choice of referring expression. Third (criterion 3), the study investigates native, adult users' production of referring expressions.[2]

Next, we provide a concise summary of the relevant studies and their findings (for an overview, see Table 6.1). Several studies have reported a positive effect of predictability on pronoun production. Among the English studies, Arnold (2001) and Weatherford and Arnold (2021) found this effect primarily for object

[1] URL: `https://scholar.google.com/`

[2] Readers may be interested to know that second language learners have also been investigated in this area. For instance, research has been conducted on Japanese- and Korean-speaking learners of English (Grüter et al., 2017), as well as Chinese-speaking English learners (Cheng and Almor, 2019).

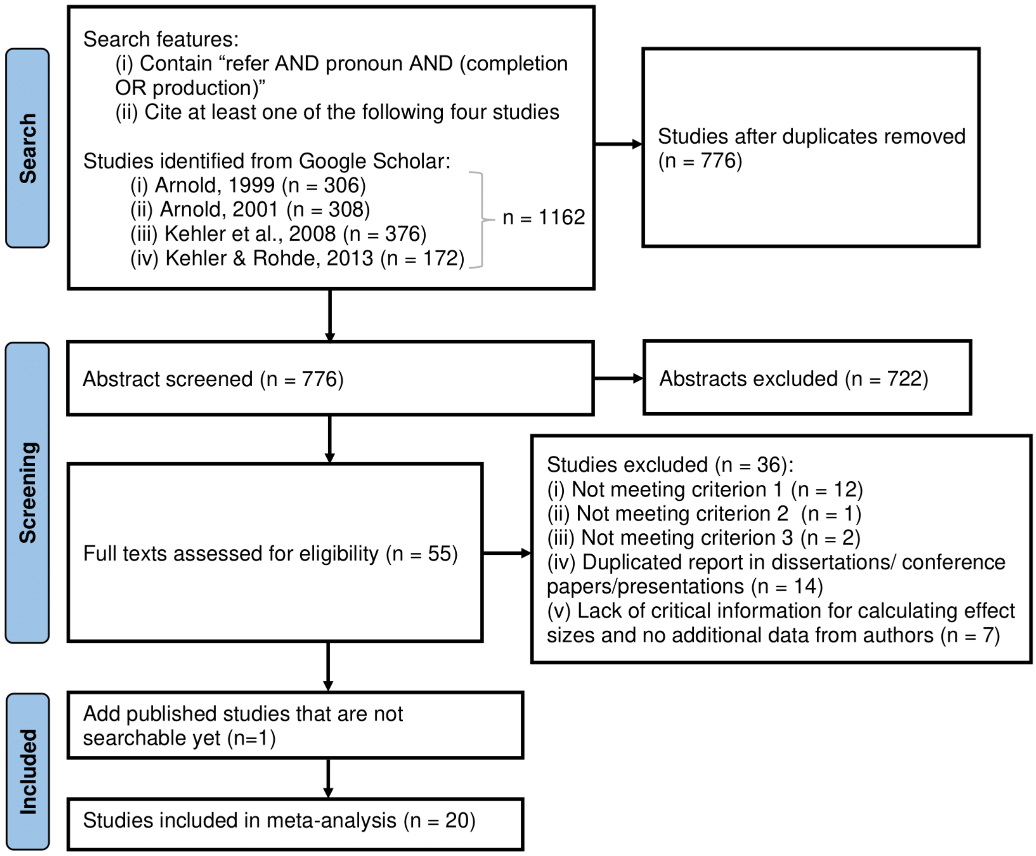

Search
Screening
Included
Search features:
(i) Contain "refer AND pronoun AND (completion OR production)"
(ii) Cite at least one of the following four studies
Studies identified from Google Scholar:
(i) Arnold, 1999 (n = 306)
(ii) Arnold, 2001 (n = 308)
(iii) Kehler et al., 2008 (n = 376)
(iv) Kehler & Rohde, 2013 (n = 172)
n = 1162
Studies after duplicates removed (n = 776)
Abstract screened (n = 776)
Abstracts excluded (n = 722)
Full texts assessed for eligibility (n = 55)
Studies excluded (n = 36):
(i) Not meeting criterion 1 (n = 12)
(ii) Not meeting criterion 2 (n = 1)
(iii) Not meeting criterion 3 (n = 2)
(iv) Duplicated report in dissertations/ conference papers/presentations (n = 14)
(v) Lack of critical information for calculating effect sizes and no additional data from authors (n = 7)
Add published studies that are not searchable yet (n=1)
Studies included in meta-analysis (n = 20)

Figure 6.1: Flow diagram showing study selection for the meta-analysis.

referents, while Rosa and Arnold (2017), Zerkle and Arnold (2019), and one of the experiments in Ye and Arnold (2023) support this conclusion more generally. Additional evidence comes from research on various languages, such as Korean (Hwang, 2023b), Spanish (Medina Fetterman et al. 2022, with effects only for overt pronouns), Romanian (Lindemann et al., 2020), and Turkish (Konuk and von Heusinger 2021, with effects only for subject referents).

In contrast, other studies have found no significant effect of predictability on pronoun use. These include our previous corpus analyses (Chapter 3), experimental study (Chapter 4),[3] Ferretti et al. (2009), Fukumura and Van Gompel (2010), Rohde and Kehler (2014), Rosa (2015), Kehler and Rohde (2019), Kravtchenko (2022), and the other experiment in Ye and Arnold (2023). Cross-linguistic support for this perspective comes from studies on Catalan (Mayol, 2018), Chinese

[3] We did not include our experimental study, as it had not undergone peer review at the time of conducting the meta-analysis.

Mandarin (Hwang et al., 2022; Zhan et al., 2020), German (Holler and Suckow, 2016), Korean (Hwang, 2023a), and American Sign Language (ASL; Frederiksen and Mayberry 2022).

Additionally, some studies have indicated a more complex pattern. Lam and Hwang (2022) found an increased use of null pronouns for less predictable referents in Mandarin. Portele and Bader (2020), on the other hand, observed lower pronoun usage for both more predictable experiencer referents and less predictable stimulus referents, suggesting that predictability cannot fully explain these data.

	study	language	manipulation	conclusion
1	Arnold (2001)	English	TPV	✓
2	Ferretti et al. (2009)[a]	English	verb aspect	✗
3	Fukumura and Van Gompel (2010)	English	ICV	✗
4	Rohde and Kehler (2014)	English	ICV	✗
5	Rosa (2015)	English	TPV	✗[b]
6	Holler and Suckow (2016)	German	ICV, relation	✗
7	Rosa and Arnold (2017)	English	TPV	✓
8	Mayol (2018)	Catalan	ICV	✗
9	Kehler and Rohde (2019)	English	relative clause	✗
10	Zerkle and Arnold (2019)	English	TPV	✓[c]
11	Lindemann et al. (2020)	Romanian	TPV	✓
12	Portele and Bader (2020)	German	relation	✗[d]
13	Zhan et al. (2020)	Mandarin	ICV	✗
14	Konuk and von Heusinger (2021)	Turkish	ICV	✓[e]
15	Weatherford and Arnold (2021)	English	ICV	✓[f]
16	Frederiksen and Mayberry (2022)	ASL	ICV	✗
17	Hwang et al. (2022)	Mandarin	ICV, TPV, relation	✗
18	Kravtchenko (2022)	English	ICV, TPV	✗
19	Lam and Hwang (2022)	Mandarin	ICV	✗[g]
20	Liao (2022)	English	relation	✗
21	Medina Fetterman et al. (2022)	Spanish	TPV	✓[h]
22	Patterson et al. (2022)	German	ICV	✗
23	Hwang (2023a)	Korean	ICV, TPV, relation	✗
24	Hwang (2023b)	Korean	connective	✓
25	Hwang and Lam (2023)	Mandarin, English	relation	✗
26	Ye and Arnold (2023)	English	ICV	✗, ✓[i]

Table 6.1: Overview of conclusions drawn in previous work.

continued

study	language	manipulation	conclusion

[a] See the same experiment also in Rohde (2008), Experiment VII.

[b] This study did not observe more pronouns produced for more predictable referents but speculated that this may have been due to an issue of power.

[c] Only the subject continuation trials were analyzed as speakers rarely used reduced expressions for non-subject continuation trials (10%).

[d] The pronominalization rate was lower for the more predictable Experiencer and the less predictable Stimulus.

[e] This study found the effect of predictability only within subjects, while there was no difference for objects.

[f] This study found the effect of predictability only within objects, while there was no difference for subjects.

[g] This study reported a negative effect of predictability: participants used more null pronouns for less predictable referents.

[h] This study found that the effect of predictability only emerged for overt pronouns when used to refer to nonsubject characters.

[i] This study reported a significant effect of predictability in a spoken experiment, whereas no such effect was observed in a written experiment.

In the screening process, dissertations and conference papers were excluded if their analyses were also reported in a peer-reviewed article. In such instances, only the peer-reviewed article was considered for inclusion (e.g., Weatherford and Arnold 2021; Zerkle and Arnold 2019). When encountering studies that lacked essential information for calculating effect sizes, we attempted to contact either the corresponding or first author to obtain unpublished data. Seven studies were ultimately excluded due to missing data or non-responsiveness from the authors. Consequently, 19 studies qualified for inclusion in the analysis. During our search, our corpus study was not yet discoverable online due to a delay in its publication. Despite this, we decided to include this study in our analysis.[4] Consequently, a total of 20 studies were incorporated into the final analysis.

Out of these 20 studies, 6 report multiple relevant experiments, each with independent samples, with slight variations in setting, or different stimuli. Specifically, Fukumura and Van Gompel (2010) conducted an experiment where one group of participants freely chose the referent, while another group was instructed to continue with a specific referent. Weatherford and Arnold (2021) carried out two experiments that only differed in the order of character mentions in the context sentence. Similarly, Medina Fetterman et al. (2022) conducted two experiments, one in written format and the other in spoken format. Hwang (2023a) manipulated

[4] The study had been published and was searchable at the beginning of 2023.

predictability by varying connectives in one experiment and verb types in another. Solstad and Bott (2022) conducted two experiments using two different prompt types, connective prompts and full-stop prompts. Contemori and Di Domenico (2021) recruited both Italian and Spanish participants to complete the task in their own language.

To explore the influence of varying experimental conditions and materials on the effect size of predictability, we included these multiple experiments as separate samples in our analysis. As a result, 20 primary peer-reviewed studies comprising data from 26 samples (no fewer than 1145 participants) were included in our meta-analysis, as summarized in Table 6.2.

	experiment	language	publication	N of participants
1	Arnold (2001)	English	journal	16
2	Fukumura and Van Gompel (2010) Pre1	English	journal	24
3	Fukumura and Van Gompel (2010) Exp2	English	journal	24
4	Rohde and Kehler (2014)	English	journal	28
5	Holler and Suckow (2016)	German	book	96
6	Mayol (2018)	Catalan	journal	78
7	Kehler and Rohde (2019)	English	journal	40
8	Zerkle and Arnold (2019)	English	journal	34
9	Portele and Bader (2020)	German	book	32
10	Zhan et al. (2020)	Mandarin	journal	50
11	Contemori and Di Domenico (2021) Exp1[a]	Spanish[b]	journal	24
12	Contemori and Di Domenico (2021) Exp2[a]	Italian	journal	24
13	Weatherford and Arnold (2021) Exp1	English	journal	56
14	Weatherford and Arnold (2021) Exp2	English	journal	46
15	Konuk and von Heusinger (2021)	Turkish	proceedings	90
16	Hwang et al. (2022)	Mandarin	journal	62
17	Lam and Hwang (2022)	Mandarin	journal	40
18	Liao (2022)[c]	English	proceedings	corpus
19	Medina Fetterman et al. (2022) Exp1	Spanish	journal	43[d]
20	Medina Fetterman et al. (2022) Exp2	Spanish	journal	26[e]
21	Patterson et al. (2022)	German	journal	40
22	Solstad and Bott (2022) Exp1[a]	German	journal	52
23	Solstad and Bott (2022) Exp2[a]	German	journal	64
24	Hwang (2023a) Exp1	Korean	journal	57
25	Hwang (2023a) Exp2	Korean	journal	65
26	Hwang (2023b)	Korean	journal	34

continued

experiment	language	publication	N of participants

Table 6.2: Samples included in the meta-analysis, with publication type and number of participants included in the analysis.

[a] These experiments did not aim at addressing our research question. However, we included them in our analysis as they employed the same experimental paradigm as other studies and their data allowed for the calculation of the effect size of predictability. We report sensitivity analyses excluding these two studies in Appendix D.4.3.

[b] This study examines Mexican Spanish, focusing on undergraduate students at Universidad Autónoma de Ciudad Juárez. The authors noted that the region's proximity to the U.S. may render the local Spanish a contact variety.

[c] Liao (2022) is the only study that investigated the question of interest through corpus analysis instead of empirical experiments. It met our criteria for inclusion and allowed for the calculation of the effect size of predictability. However, it did not control for contextual features such as referent animacy, verb aspect, or the number of referents, which may potentially introduce variability in the findings. In order to rule out potential biases, we conducted a sensitivity analysis excluding our corpus study (see Appendix D.4.1).

[d] Participants were from Argentina, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala, Mexico, Panama, Peru, Puerto Rico, Spain, Uruguay, and Venezuela.

[e] The study involved participants from seven countries/territories, residing in the United States for 0-21 years, with one individual raised between Colombia and Argentina. 19 participants were multilingual, speaking at least one additional language.

6.2.3. Effect sizes: Odds ratios

In this meta-analysis, effect sizes were recorded as *odds ratios*. The odds ratio is a statistical measure that evaluates the relationship between two properties in a population. It is frequently used when there are comparison pairs (pairs of groups that are compared against each other) and when the variable of interest is dichotomous, like in our case where referents with higher predictability are contrasted with those with lower predictability and the use of pronouns versus other referential forms is the variable of interest.

An example of how odds ratios are calculated is given in Equation 6.1, on the basis of the made-up data in Table 6.3. The resulting odds ratio of 2.67 suggests that pronouns are 2.67 times more likely to be used with the more predictable referent. An odds ratio of 1 would instead suggest no effect of predictability on

	more predictable referent	less predictable referent
pronoun	40 (A)	30 (B)
non-pronoun	10 (C)	20 (D)

Table 6.3: Example contingency table with made-up data: Number of referring expressions produced for the more predictable referent (A and C) and the less predictable one (B and D).

pronoun production; and an odds ratio smaller than 1 suggests a negative effect of predictability.

$$\text{odds ratio} = \frac{A \times D}{B \times C} = \frac{40 \times 20}{30 \times 10} = 2.67 \tag{6.1}$$

In analogy to Table 6.3, for each experiment, we gathered the number of continuations and the pronominalization rate in each condition. In doing so, the same sample sometimes contributes multiple odds ratios. This happens under three different circumstances. First, when an experiment examined predictability while controlling for grammatical roles. For instance, production data for minimal pairs constructed using Goal-Source verbs and Source-Goal verbs contribute two odds ratios: one that compares Goal-Subject to Source-Subject, and another that contrasts Goal-Object with Source-Object. The majority of studies included in our analysis exhibit this characteristic (Arnold, 2001; Hwang et al., 2022; Hwang, 2023a; Medina Fetterman et al., 2022; Zerkle and Arnold, 2019; Contemori and Di Domenico, 2021; Fukumura and Van Gompel, 2010; Holler and Suckow, 2016; Hwang, 2023a; Konuk and von Heusinger, 2021; Mayol, 2018; Patterson et al., 2022; Rohde and Kehler, 2014; Solstad and Bott, 2022; Weatherford and Arnold, 2021; Zhan et al., 2020; Kehler and Rohde, 2019; Portele and Bader, 2020). Second, it is also the case when a single sample is on pro-drop languages, measuring production rates of both null and overt pronouns. This kind of study produces two effect sizes, one for overt, one for null pronouns. An example is Medina Fetterman et al. (2022), who finds that predictability primarily affects the use of Spanish overt pronouns but not null pronouns. The same holds for Catalan (Mayol, 2018), Mandarin (Zhan et al., 2020; Hwang et al., 2022; Lam and Hwang, 2022), Turkish (Konuk and von Heusinger, 2021), Italian and Spanish (Contemori and Di Domenico, 2021). Finally, a single sample may contribute multiple odds ratios when the experiment combines various types of stimuli or manipulates predictability in multiple ways (Holler and Suckow 2016; Hwang et al. 2022; Experiment 1 & 2 in Hwang 2023a; Experiment 1 in Solstad and

Bott 2022). For instance, Hwang et al. 2022 conducted an experiment using both transfer-of-possession verbs (Source-Goal & Goal-Source) and implicit causality verbs (Experiencer-Stimulus & Stimulus-Experiencer). Analyzing the results within each verb category can generate at least one odds ratio, with the possibility of deriving additional odds ratios when also considering the previous two circumstances (e.g., four odds ratios for each verb category: production of null pronouns for subject referents, overt pronouns for subject referents, null pronouns for object referents, overt pronouns for object referents).

After teasing apart samples according to these three cases, we end up with a total of 104 effect sizes that speak to referent predictability. They draw from 20 peer-reviewed articles, comprised of 26 individual experiments/samples. The fact that some effect sizes draw from the same study or sample is reflected in the multi-level model structures we employ (see section 6.2.4).

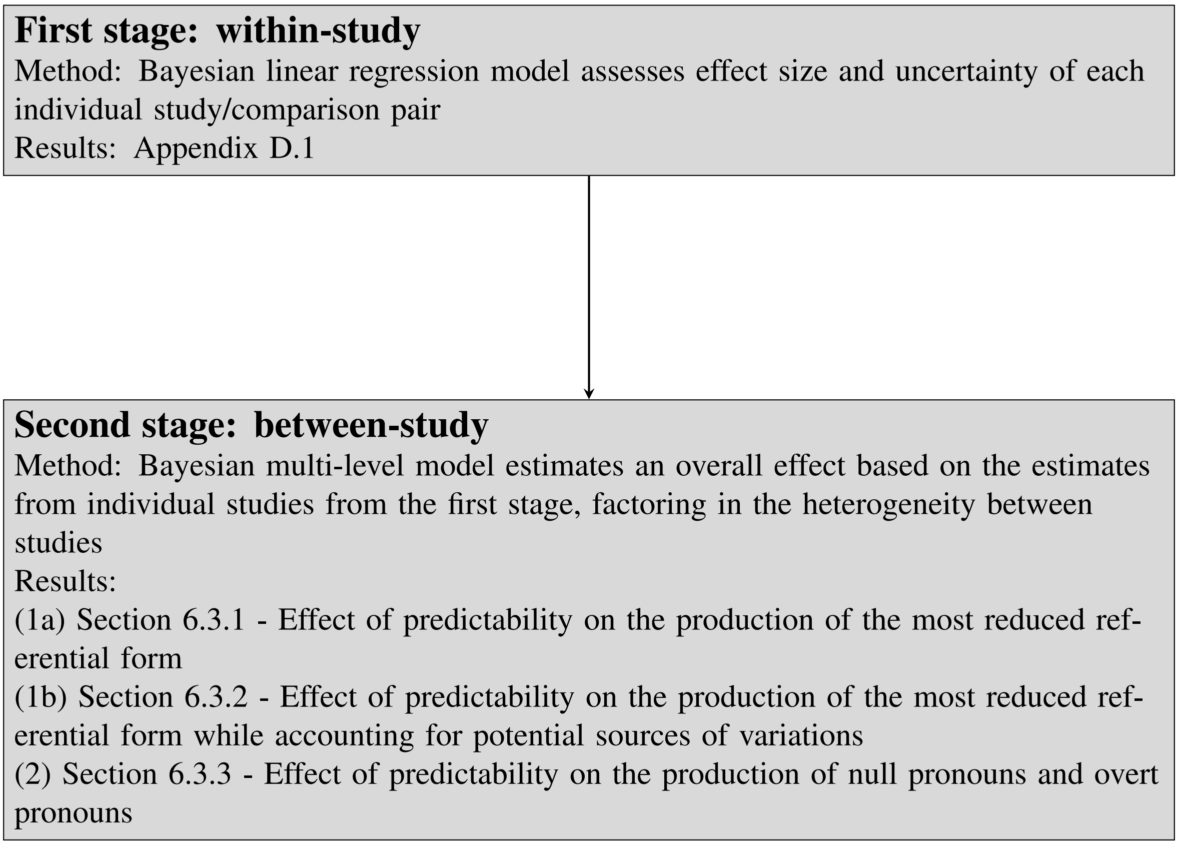
Note that our meta-analytical models estimate the effect size on the scale of log odds (as we explain in Section 6.2.4). However, we transform them into odds ratios when presenting the effect sizes for ease of interpretation.

6.2.4. Analyses

All analyses were performed using Bayesian inference methods, using the *brms*-package (Bürkner, 2021) of *R* (version 4.1.2, R Core Team 2021). All fits were run with four chains for 2000 iterations each, with half as warm-up. The models use Student's-t priors (df = 3, μ=0, and σ=2.5). All fits were diagnosed to rule out pathologies in their estimates. All had parameters with a split $\hat{R} < 1.1$ (Gelman and Rubin, 1992), suggesting well-mixed chains; they had no saturated trajectory lengths (i.e., the sampler did not stop prematurely); they had no divergent transitions (i.e., no difficulties in exploring the posterior); and they all had an energy Bayesian Fraction of Missing Information over 0.2 (i.e., no inefficiency in the momentum resampling between trajectories (Betancourt, 2017).

The meta-analysis proceeds in two stages (Higgins et al., 2019), as depicted in Figure 6.2.

First, a within-studies stage assesses effect size and uncertainty of each individual study / comparison pair, arriving at a unique effect size estimate per study / comparison pair. A second, between-studies, stage then infers a single grand pooled effect size, based on the individual estimates from the first stage. The intuition behind this procedure is that the findings of each individual study are a draw from a distribution of effect sizes particular to that study, with an (unobserved) true effect underlying it. This corresponds to stage one. Behind each true effect particular to individual studies, however, meta-analyses assume an overall (again, unobserved) effect distribution from which they all draw. This corresponds to stage two.



First stage: within-study

Method: Bayesian linear regression model assesses effect size and uncertainty of each individual study/comparison pair

Results: Appendix D.1

Second stage: between-study

Method: Bayesian multi-level model estimates an overall effect based on the estimates from individual studies from the first stage, factoring in the heterogeneity between studies

Results:

(1a) Section 6.3.1 - Effect of predictability on the production of the most reduced referential form

(1b) Section 6.3.2 - Effect of predictability on the production of the most reduced referential form while accounting for potential sources of variations

(2) Section 6.3.3 - Effect of predictability on the production of null pronouns and overt pronouns

Figure 6.2: Diagram illustrating the two stages of the analysis.

More precisely, the first stage here consisted in estimating the effect of each individual study with regard to pronoun use predicted by referent predictability using linear regression models. The dependent variable *pronoun use* was coded to be 1 whenever a referent is realized using a pronoun, and 0 otherwise. We then coded re-mentions of the more predictable referents (e.g., stimulus referent "Alan" in "Paul liked Alan because Alan...") as 1 and re-mentions of the less predictable referents (e.g., experiencer referent "Alan" in "Paul embarrassed Alan because Alan...") as 0. By fitting linear models, we extracted posterior distributions of the estimate of the effect, which represents the difference in pronominalization between referents that were more predictable and those that were less predictable in each comparison pair.

As mentioned above, in the second stage, we estimate an overall effect based on the estimates from individual studies from the first stage. To do this, we performed multi-level meta-analyses that factor in the heterogeneity between studies. Because, as explained above, some articles reported results from multiple samples, population-level effects were included not only at the article level but also at the within-article level.[5]

We carry out two separate meta-analyses. The goal of the first one is to quantify the effect of predictability on the production of the most reduced referential form available in each language. This minimal form varies between Germanic languages, like English or German, where it is an overtly expressed pronoun, and pro-drop languages, such as Catalan or Mandarin, where the most reduced form available is a null pronoun.

For this first meta-analysis, we constructed two separate models. The first model assumes that all included studies are comparable and that there are no important characteristics that distinguish them. In contrast, the second model, informed by the literature, identifies three potential moderators, i.e., variables that may influence the relationship between the independent variable *referent predictability* and the dependent variable *pronoun use*. These variables are:

1. Manipulation of referent predictability: implicit causality verbs, transfer-of-possession verbs, discourse relations, and relative clauses. Previous research has speculated that the influence of predictability may be context-dependent, with its effects being more pronounced in specific contexts, such as transfer-of-possession, while being more difficult to discern in others, such as implicit causality (e.g., Rosa and Arnold, 2017).

[5] In *brms* syntax, this corresponds to "formula $=$ estimate $|$ se(error) $\sim$ $1 + (1$ $|$ article/sample)". The fixed intercept item 1 represents the estimated average mean over studies, and the intercept (1 | article / sample) allows for estimating the heterogeneity between studies and between samples nested within articles.

2. Language family: Romance (Catalan, Italian, Spanish), Mandarin, Korean, Turkish, and Germanic (English, German). Pronouns exhibit varying behavior across different languages; for example, within null-subject languages, Mandarin and Romance null pronouns do not necessarily function in the same manner (e.g., Zhan et al. 2020; Filiaci et al. 2014). Consequently, the impact of predictability on pronoun usage may differ across language families[6].

3. Grammatical role: subject or object. Some studies have found that the effect of predictability was stronger for references to the object than for references to the subject, arguing that this may be due to the overall high use of pronouns for subjects (e.g., Weatherford and Arnold 2021).

Our second model factors in the influence of these three potential moderators by including them as predictors in the models. To facilitate interpretation of the results, all predictors were centered around their respective means. Consequently, coefficients in the models represent the deviations of each predictor from its mean value.

We note that a recent study by Ye and Arnold (2023) also identified a distinction in task modality (written vs. spoken tasks) in relation to referent predictability and pronoun usage. While the written task exhibited no effect of predictability on pronoun usage, the spoken task demonstrated a significant impact, suggesting that the influence of predictability was more pronounced in communicative environments involving direct addressees. Task modality was not included as a covariate in our analyses due to data limitations. Specifically, the majority of available data were collected from written tasks, and the scarce spoken data were primarily in English (3 out of a total of 4 spoken studies). Moreover, our analysis also included our previous corpus-based study (see Chapter 3), which used both written and spoken corpus data. As a result, we would have needed to introduce a third level, "corpus-based", to the task modality factor; however, only one study provided such observations. Considering the data imbalance, accurately estimating model parameters would be challenging if we were to include task modality as a factor in our analysis. Therefore, we instead took this factor into account in a post-hoc exploratory analysis of the impact of task modality on the relationship between referent predictability and pronoun use using a subset of the dataset (see Appendix D.5 for the results and a more detailed discussion).

[6] In the analyses, languages were classified according to their respective language families, primarily due to the limited availability of data for individual languages. To ensure the validity of our subsequent results, we performed a sensitivity analysis, using Language as a covariate instead of Language family. As expected, using Language results in greater uncertainty concerning the summary effect estimate, with the general pattern of the estimates unchanged. We report this analysis in Appendix D.4.4.

In the second meta-analysis, we focused specifically on pro-drop languages: Catalan, Italian, Korean, Mandarin, Spanish, and Turkish. These languages permit null subjects and consequently speakers can choose between two types of pronouns (null and overt pronouns). We consider them both in this analysis to assess potential differences in the impact of predictability. Indeed, previous studies have suggested that null and overt pronouns may be constrained by different factors and to varying degrees (Filiaci et al., 2014; Fedele and Kaiser, 2015). For instance, Medina Fetterman et al. (2022) found that in Spanish predictability primarily affected overt pronoun production but not null pronoun production. Overt pronouns, though more explicit than null forms, still represent a reduced referential form compared to names or full noun phrases. Pro-drop languages thus provide a valuable context for investigating the consistency of predictability effects on the production of pronominal forms with varying degrees of reduction. This can contribute to our understanding of the semantic constraints governing the usage of diverse pronominal forms, a crucial aspect for developing reference production models in null-pronoun languages, as underscored by Medina Fetterman et al. (2022). Furthermore, given that previous research has largely focused on English, examining referential choices in languages with a broader range of referring expression types may reveal patterns that would otherwise remain undiscovered (Vogels, 2019).

To assess variations in predictability effects between null and overt pronouns, we incorporated pronoun type as an additional predictor in this analysis. We also controlled for the three factors considered in the previous model: language family, grammatical role, and manipulation type.

The individual effect sizes estimated during the first stage (within-study) can be found in Appendix D.1. In the following results section, our primary focus will be on the outcome of the multi-level level meta-analyses conducted during the second stage (between-study), as these synthesize the results from multiple studies and directly address our main research question.

6.3. Results

6.3.1. Effect of predictability on the use of the most reduced reference form

We first fit a basic population-level model with no individual-level predictors. That is, this model only factors in variation across studies and within samples when estimating the overall effect of predictability across studies (26 independent experiments comprising 73 odds ratios; see Section 6.2.3 for the cases where a single experiment contributes multiple odds ratios). This yields an overall estimated

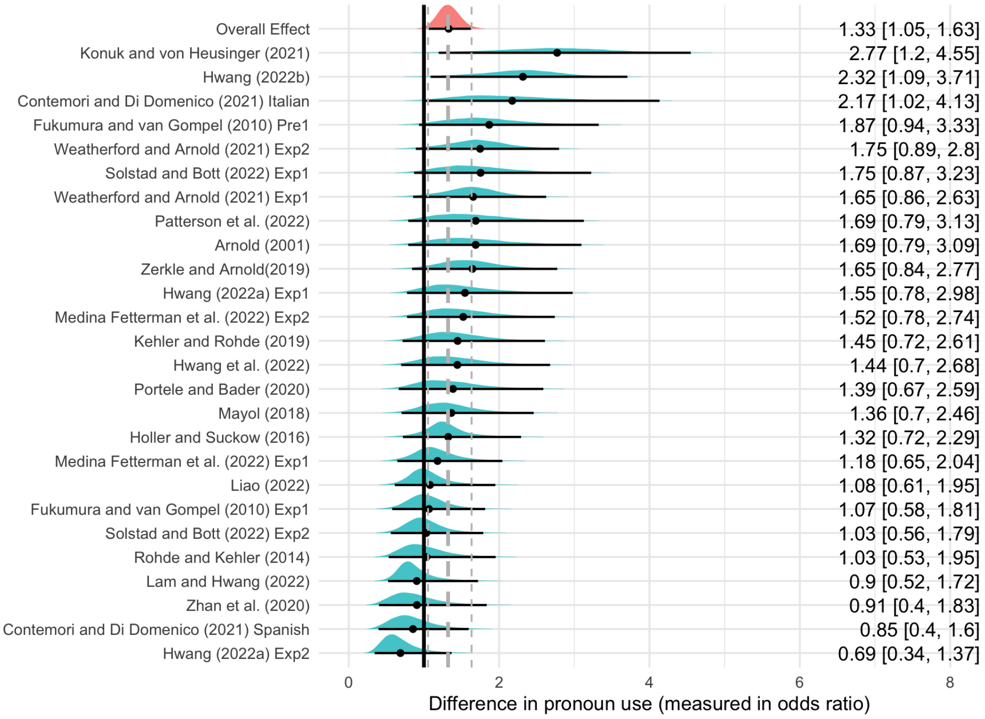


Figure 6.3: Forest plot of the estimates of the difference in the use of the most reduced referential form between the more predictable referents and the less predictable referents.

odds-ratio of 1.33 [1.05, 1.63, 95% CIs], suggesting that the most reduced referential form is around 1.33 times more likely to be used for the more predictable referents than for the less predictable referents. The 95% credible interval ranges from 1.05 to 1.63, suggesting that there is an effect of predictability on the production of the most reduced referential form, albeit small. Effect sizes of this magnitude, which suggest a potentially subtle difference, can be challenging to identify.

The overall effect, together with the estimated odds ratios of the individual experiments, is shown in Figure 6.3. The top-most row corresponds to the overall estimate obtained from collating the evidence from the individual studies below it. Note that these estimates do not represent the estimates of the individual studies, obtained in the first stage of our meta-analysis (See Table D.1 and D.2 in Appendix D.1 for these estimates). Instead, they are re-adjusted estimates of individual studies, once the data coming from other studies is considered. This allows the results to draw strength from the data of other studies.

We also conducted a sensitivity analysis that excluded our corpus study, reported in Appendix D.4.1. This is because our corpus study used re-mention

	Est. mean	Est. error	95% CI
Intercept	0.43	0.22	[0.01, 0.86]
manipulation ICV	-0.09	0.18	[-0.45, 0.26]
manipulation relativeClause	-0.03	0.44	[-0.92, 0.85]
manipulation relation	-0.10	0.20	[-0.51, 0.26]
languageFamily Turkic	0.92	0.46	[0.04, 1.87]
languageFamily Mandarin	-0.51	0.31	[-1.13, 0.09]
languageFamily Korean	-0.10	0.31	[-0.71, 0.52]
languageFamily Germanic	-0.07	0.20	[-0.47, 0.32]
subjectOrNot object	0.04	0.04	[-0.05, 0.12]

Table 6.4: Effect of predictability on the use of the most reduced reference form: Summary of the model with three predictors.

statistics from corpora, unlike the others, which used controlled experiments using story continuation tasks with lab-constructed stimuli. The assumption in our corpus study is that these statistics hearers may track statistical regularities in their input to predict upcoming information, and thus corpus data could capture the distributional patterns used to estimate predictability. However, corpus contexts lack strict control for features such as referent animacy. The results of the sensitivity analysis suggest that the inclusion of the corpus-based study introduced no bias to the meta-analysis.

6.3.2. Effect of predictability on the use of the most reduced reference form while controlling for potential sources of variations

This section presents the analysis in which we incorporate three predictors to account for potential variation among the individual samples: grammatical role; manipulation type; and language family; (see Section 6.2.4). After adjusting for these factors, the effect estimate increases to 1.54 [1.01, 2.36, 95% CIs]. However, the addition of further parameters to the model, when the available data may be insufficient for proper estimation, decreases the confidence in the overall effect estimate of predictability (note the wide 95% CI of [1.01, 2.36]). The summary of the model is given in Table 6.4. When it comes to the three predictors, this analysis does not provide strong evidence for their influence on the use of the most reduced reference form. We discuss this matter in more detail in Appendix D.2.

6.3.3. Effect of predictability on the production of null pronouns and overt pronouns

The third analysis narrows down the meta-analysis to focus on pro-drop languages only. The dataset for this focused analysis contains data from 6 languages (Italian, Spanish, Mandarin, Korean, Turkish, and Catalan), obtained from 12 independent samples across 9 distinct studies. As before, the three potential sources of variation are included as predictors (Section 6.2.4), and we add a new predictor that codifies the different pronominal forms (null and overt pronouns). Additionally, we incorporated the interaction between pronoun type and grammatical role to account for potential variations in predictability effects on overt and null pronoun usage, based on the antecedent's grammatical role.

This interaction is informed by existing research. It has been posited that null pronouns typically refer to prominent antecedents, while overt pronouns are more likely to reference less prominent ones (e.g., Ariel, 1990). The prominence of a referent is influenced by its grammatical role; subjects, which are generally more prominent, are frequently expressed by null pronouns. In contrast, less prominent referents, such as objects, tend to be associated with more explicit expressions like overt pronouns (e.g., Mayol, 2018). This distinction is supported by empirical evidence on comprehension in languages like Spanish, where readers experience slower processing times when an overt pronoun refers to the prior subject, as opposed to sentences with a null pronoun (e.g. Gelormini-Lezama and Almor, 2011, 2014; Gelormini-Lezama, 2018). Consequently, if predictability indeed plays a role, one would expect a higher prevalence of null pronouns for more predictable and prominent subject referents, while an increased overt pronoun usage for these subject referents is less expected.

For object referents, however, there were fewer occurrences of null pronouns compared to overt pronouns. This distribution may complicate the detection of the effect on the production of null pronouns for object referents, as suggested in previous research (e.g., Lam and Hwang 2022; Hwang et al. 2022). In contrast, an effect of predictability on the production of overt pronouns is more likely to emerge for object referents. This is supported by Medina Fetterman et al. (2022), who consistently found an increased usage of overt Spanish pronouns for more predictable object referents across experiments, but not for more predictable subject referents.

The model summary is presented in Table 6.5.[7] Focusing on pro-drop lan-

[7] We excluded Korean overt pronouns from our analysis due to their rarity and similarity to noun phrases (Kim, 1990; Choi, 2013; Hwang, 2023a). Retaining Korean data for null pronouns poses no harm, as the data is partially pooled across languages, contributing evidence solely for the null-subject instances; however, caution is advised when interpreting language-level predictors for Korean.

	Est. mean	Est. error	95% CI
Intercept	0.47	0.24	[0.01, 0.97]
manipulation ICV	-0.37	0.15	[-0.67, -0.07]
manipulation relation	0.21	0.14	[-0.06, 0.47]
languageFamily Turkic	0.74	0.51	[-0.27, 1.79]
languageFamily Mandarin	-0.26	0.36	[-0.97, 0.52]
languageFamily Korean	-0.34	0.40	[-1.15, 0.47]
subjectOrNot object	0.07	0.07	[-0.06, 0.20]
pronounType overt	0.01	0.06	[-0.12, 0.14]
subjectOrNot object:pronounTypeovert	0.15	0.06	[0.03, 0.26]

Table 6.5: Effect of predictability in pro-drop languages: summary of the model with four covariates added.

guages only, we obtain an effect estimate of 1.60 [1.01, 2.64, 95% CIs].[8] This is larger yet more uncertain than the previous estimate that included Germanic languages. Once more, we find no clear evidence supporting a major influence of grammatical role, manipulation type, or language family on the results. The observed trends for each of these variables roughly align with the findings of the second analysis presented in Table 6.4, indicating that the direction in which each moderator influences the relationship between referent predictability and pronoun use is consistent across analyses. We discuss this matter in more detail in Appendix D.3. All in all, overall findings are again that the effect of predictability on pronoun production is present but small, and the examined factors do not prominently moderate this effect.

One concern is the comparability of our Spanish data with data from other Romance languages included in the analyses (Italian and Catalan). We obtained Spanish data from two independent studies, Contemori and Di Domenico (2021) and Medina Fetterman et al. (2022). As noted in Table 6.2, the variety of Spanish tested in Contemori and Di Domenico (2021) may be considered a contact variety, and participants in Medina Fetterman et al. (2022) were from different countries/territories and had varying lengths of residence in the United States. We performed a sensitivity analysis, excluding the Spanish experiments, to check the possibility that this introduces noise into the data due to possible variations in overt and null subject pronoun usage across different varieties of Spanish (e.g., Alfaraz, 2015). The results indicate that the inclusion of Spanish data does not qualitatively impact or distort our results. For more details, see Appendix D.4.2.

[8] The odds ratio of 1.60 is computed by exponentiating the coefficient 0.47 (in log odds) presented in Table 6.5.

6.4. Discussion

To address our research question in a systematic way, we conducted a meta-analysis synthesizing results from 20 independent studies, which comprise data from 26 experiments and 8 languages. Our primary objective was to investigate the effect of predictability on the production of the most reduced reference form, specifically pronouns in Germanic languages (English and German), as well as null pronouns in the case of pro-drop languages (Catalan, Italian, Mandarin, Korean, Spanish, and Turkish).

Overall results Our meta-analysis suggests that there is indeed an effect of referent predictability on the production of the most reduced reference form. That being said, it is likely not a particularly large effect. Specifically, the estimated overall effect is an odds ratio of 1.33 ([1.05, 1.63], with 95% CIs), indicating that the odds of using pronouns for more predictable referents are moderately higher than the odds of using pronouns for less predictable referents. The estimated overall effect remains comparable in magnitude, although with heightened uncertainty (odds ratio of 1.54 with 95% CIs of [1.01, 2.36]), when accounting for potential sources of variation from grammatical roles (subjects or objects), manipulation type (transfer-of-possession verbs, implicit causality verbs, discourse relations, and relative clauses), and language family (Germanic, Korean, Mandarin, Romance, Turkish). All in all, these results suggest a small to modest positive effect of referent predictability on pronoun production.

Our second analysis specifically targets pro-drop languages, synthesizing data from 9 studies, comprising 12 experiments and 6 languages. This analysis also suggests a small to modest overall effect of predictability, again with notable uncertainty about the true effect (an odds ratio of 1.60 with 95% CIs of [1.01, 2.64]).

Thus, the overall effect estimated in our meta-analysis supports the Expectancy Hypothesis (e.g., Arnold, 2001, 2010), that is, the hypothesis that referent predictability does influence pronoun production. According to this framework, as the predictability of referents increases, the referents become more salient in addressees' mental representation of discourse. As a result, speakers are more inclined to use more reduced forms for these referents, thereby signaling the addressees to retrieve the readily accessible referents from memory. However, the extent to which predictability (at least, the coherence-driven predictability investigated in this meta-analysis) contributes to enhanced accessibility may be limited, in the sense that the true effect seems to be relatively small. This may actually be taken to challenge the Expectancy Hypothesis, which considers predictability/next-mention expectation to be a strong influencing factor on the accessibility or activation of a referent. Robust evidence indicates that, instead, structural and grammatical cues, such as grammatical role, exert a more substantial impact on pronoun production (e.g., Rohde and Kehler, 2014; Fukumura and Van Gompel, 2010; Medina Fetterman et al., 2022).

One possible explanation for this difference is a trade-off in cognitive effort vs. communicative efficiency for speakers. Both predictability and grammatico-structural factors are useful to signal the intended referent to the addressee; but relying on predictability arguably imposes a greater cognitive load on speakers, as they need to continuously assess and update the discourse model to determine the most predictable referent for their addressees. In comparison, grammatical and structural factors offer a more stable foundation for speakers, enabling faster and less cognitively demanding choices in selecting referring expressions: These factors need to be tracked for other reasons, such as determining which inflectional form to use for subject-verb agreement. Therefore, they come at little to no additional cognitive load. In line with the small to modest effect that we find, speakers may utilize predictability to a limited extent, balancing its costs and benefits during language production and comprehension.

Limitations While our meta-analysis suggests that there is evidence for a small to modest positive effect of predictability on pronoun production, we should also stress that the case is by no means closed. The current meta-analysis, which is based on a limited sample of 20 studies, contains some uncertainty in its estimates. Future work can inform new meta-analyses that build on the present one.

It is also important to note that, we have made a conscious decision not to incorporate two potentially influential factors in this meta-analysis: character gender (same or different; Kravtchenko 2022; Medina Fetterman et al. 2022) and constraints on narrative continuation (whether participants were required to continue with one character or allowed the liberty to select the subsequent mention; Kravtchenko 2022). Our rationale for this exclusion stems from the fact that including these variables would introduce additional levels to the analysis, and current available data is insufficient to accurately estimate these levels. Incorporating these factors would consequently lead to heightened uncertainty in our findings. We encourage future research to explore these factors further when more extensive data becomes available.

Recommendations for future research To start addressing the heterogeneity present in the current research landscape, we examined the influence of four distinct variables on the relationship between predictability and pronominalization, namely grammatical role, language family, manipulation type, and pronominal type. While our study cannot conclusively attribute the divergent findings to these factors due to insufficient evidence, it identifies specific aspects where evidence is lacking, helping define the most promising paths to pursue.

First, it is important for future research to consider potential differences across language families and conduct more typologically diverse experiments. In our analysis, evidence of cross-linguistic variation primarily comes from Turkish in the study of Konuk and von Heusinger (2021) (see Appendix D.2), who reported an 86% null pronoun production rate for highly predictable subject antecedents, in contrast to a 49% rate for their less predictable counterparts. This 37% disparity between the two pronominalization rates in Turkish is noteworthy when compared to other language families, where the average difference is a mere 5.5%. Given that the evidence comes from a single study, it is necessary to conduct further experiments to assess these differences. We recommend further studies of the relationship between predictability and pronoun usage in Turkic languages and other language families not yet explored in the literature.

Second, we recommend that future research adopt experimental conditions that avoid eliciting pronominalization rates that fall near the bottom (approximately 0%) or the ceiling (approximately 100%). Such rates can limit the detection of the effects of referent predictability on pronoun production, as follows. In our analysis, we observe that in pro-drop languages, the production rate of null pronouns in object referents is often close to 0, particularly in some experiments involving languages such as Mandarin (Hwang et al., 2022) and Korean (Hwang, 2023a). For example, in an experiment conducted by Hwang et al. (2022) on Mandarin Chinese, the null pronoun production rate is 0.36% for more predictable object referents and 0.98% for less predictable object referents. This near-zero production rate of null pronouns for object referents substantially limits the potential to observe any variations due to predictability: If the null pronoun usage approaches the bottom, there is no room for further reduction due to predictability. Analogously, when pronominalization rates approach the ceiling, it becomes challenging to observe any additional increase in pronoun usage as a result of predictability. This issue can be particularly pronounced in studies employing crowdsourced subjects, such as those from Amazon Mechanical Turk, since crowdsourced subjects are often paid per task, creating an incentive for them to complete tasks as quickly as possible. To optimize their task completion rate, workers may favor the use of pronouns, which are less explicit but more time-efficient than other more time-consuming referring expressions. To address this issue, we recommend future research to conduct experiments in controlled lab settings or recruit participants from different sources. This approach can help reduce potential biases introduced by crowdsourcing. Future research could also explore alternative experimental conditions. For instance, most previous studies used narrative language in their experimental stimuli, which tends to elicit more frequent use of pronouns (especially for referring back to subject antecedents), potentially resulting in ceiling effects. Using alternative contexts, such as descriptive language, may yield a lower rate of pronoun use in general, providing a more varied use of referring expressions. Other methods adopted by previous research include selecting participants who exhibit variation in their expressions (see Rosa and Arnold 2017 for further discussion).

Moreover, future work is necessary to explore which specific experimental conditions make the effect of predictability more or less likely to arise. In particular, it could be that the predictability effect is more salient in contexts involving transfer-of-possession verbs compared to implicit causality verbs, as hinted at by the weak evidence in our analysis (see Appendix D.3). Rosa and Arnold (2017) highlighted several distinctions between the two verb types, as follows. First, implicit causality verbs (experiencer/stimulus verbs) like "admire" depict a feeling or a mental state and are considered atelic, meaning they lack an inherent endpoint. In contrast, transfer-of-possession verbs like "give" are telic, entailing a clear endpoint. The presence of an endpoint in transfer verbs may facilitate the conceptualization of events, making it cognitively simpler to understand and organize information related to them. This, in turn, could give rise to a more robust discourse model, characterized by a stronger and more stable mental representation during the process of language comprehension. Second, the coherence relations that support the goal effect are those that advance the narrative or outline the consequences of the initial event. Therefore, the chronological sequence of continuations mirrors the chronological order of events. In contrast, implicit causality effects primarily emerge in explanations and rely on pre-event information about the cause, which could be more difficult to access. Third, it is plausible that the experiencer, despite not being the implicit cause, holds particular prominence in the discourse. This prominence could be attributed to the fact that implicit causality verbs convey the experiencer's mental state or feeling. By emphasizing the experiencer's perspective, these verbs may inherently draw attention to the experiencer's role in the unfolding narrative. In light of all these distinctions, it is plausible for the predictability effect to be more salient in transfer-of-possession verbs, and future work should test this possibility.

Another implication of our study is the need for an adequately large sample size, as well as the importance of controlling for other more influential factors when investigating the impact of coherence-driven predictability on pronoun production. Recall that our analysis suggests that the effect size of predictability is small to modest. The magnitude of the effect size is a key factor in determining the required sample size for an informative study. With noise or variability in the data, smaller effect sizes are more challenging to detect. For studies on the role of predictability, thus, a larger sample size than is usual in the field is needed. Additionally, this small effect size of predictability implies that its influence might be overshadowed by more potent factors, such as the well-established effects of grammatical roles, as noted by Vogels (2019). This highlights the importance of controlling for these factors when investigating coherence-driven predictability in pronoun production.

Another important avenue for future research is to examine the role of individual differences in the predictability effect. If tracking predictability imposes an increased cognitive burden on speakers, those with greater cognitive resources, such as higher working memory capacity, may be better at tracking the listeners' predictability and using this information to plan and produce upcoming utterances. Such speakers may be better able to maintain a more detailed and updated representation of the discourse, allowing them to more effectively exploit predictability cues during language production.

In a similar vein, the cognitive demands for the two modalities of language production in our exploratory analysis (see Appendix D.5)–speaking and writing–may also differ. Speaking could arguably put fewer cognitive burdens given that it is faster, which requires ideas to be stored in memory for shorter periods before expression. Additionally, the presence of an interlocutor can enhance motivation to tailor expressions to the listener. In contrast, writing demands more physical energy, is usually acquired later in life, practiced less frequently, and necessitates the activation of graphemic representations for accurate spelling. As a result, subjects may be more likely to exploit predictability cues in spoken settings, where cognitive demands are lower, rather than written ones.

Finally, we strongly advocate for the adoption of open science practices in linguistics. While open science practices also encourage transparency, bolster credibility, and enhance the replicability of research findings, their role in facilitating collaboration is particularly noteworthy. By making data and materials accessible, researchers enable others to build upon existing work, combine datasets, and conduct more extensive and powerful analyses, such as the present meta-analysis. This collaborative approach creates a supportive environment that nurtures new insights and allows for the identification of patterns that may have eluded detection in smaller, individual studies.

Our study, being the first meta-analysis on the topic of referent predictability and pronoun usage, carries both methodological and theoretical implications for research on this topic and linguistics studies more broadly. By integrating and synthesizing findings across various independent studies, meta-analyses offer a powerful tool for identifying overarching patterns and addressing inconsistencies in the literature. Tacking stock, our synthesis of the present empirical landscape suggests that the effect of coherence-driven predictability on pronoun production is likely positive and small to moderate. This implies that speakers use predictability as a cue, which listeners are sensitive to, at least to some degree. Additional research is required to explore potential variations in the observed effect under diverse conditions and across diverse contexts. In broader terms, we hope that this study can also serve as an example for future studies to build on. The resources used in this analysis will enable future researchers to pose and address questions of their own design, fostering continued meta-analysis and data aggregation in the areas of predictability and pronominalization.

Chapter 7

CONCLUSIONS

This thesis investigates whether addressees' expectations about a referent being mentioned next, which is referred to as referent predictability, influence speakers' decisions regarding the pronominalization of this referent. The relationship between the two is crucial to understanding the underlying mechanisms that facilitate coordination between speakers and addressees. However, despite good theoretical grounds linking the two, the empirical evidence has so far yielded an inconclusive outcome.

Previous studies on this topic primarily utilized psycholinguistic passage continuation experiments with targeted materials featuring specific verb types as stimuli. In this thesis, I have utilized multiple methods, each with its own novelties, to explore this question. In chapters 3 and 4, we extracted passages of three discourse relations (Narration, Contrast, and Result) from richly-annotated corpora developed in the computational linguistic research. We asked how predictability, primarily induced by discourse relations, influences pronoun usage across corpus texts, and in a traditional continuation experiment but with more naturalistic contexts as experimental stimuli. In Chapter 5, we explored a more scalable approach to investigate this question, without the constraints posed by limited annotated data and the expenses associated with human participation. Specifically, we used computational estimates of referent predictability from a neural network model as proxies for human predictions. We investigated whether the referential expectation biases exhibited by the computational model are in line with existing evidence on human behavior. If so, what is the relationship between referent predictability and the choice of pronominalization if we use computational estimates of the former? Lastly, Chapter 6 reports a meta-analysis, in which we quantitatively synthesized the results of 20 independent studies on this topic, including data from over 1,145 unique participants. We aimed to understand what the overall picture looks like when we had an increased statistical power and under which conditions a predictability effect on pronominalization could be found. Throughout the thesis, I have operationalized referent predictability in varied ways. In the corpus analyses (Chapter 3), it is measured by the re-mention frequency of a referent across the corpus texts. In the completion experiment (Chapter 4), we approximate it by observing next-mention biases in participants' passage continuations. Meanwhile, in the computational modeling study (Chapter 5), referent predictability is defined using information-theoretical concepts like surprisal and entropy.

The contributions of this thesis are twofold: theoretical and methodological. On the theoretical front, the research in this thesis offers insights into theories and models pertaining to the predictive nature of human processing and pronoun production. It sheds light on the pragmatic mechanisms that facilitate coordination between speakers and addressees during discourse. On the methodological side, the thesis introduces novel, primarily computational, approaches to an open question in theoretical linguistics. For each study presented in this thesis, I have discussed its implications from both the theoretical and methodological viewpoints in the concluding sections of their respective chapters. In this final chapter, I discuss the divergent outcomes observed across studies and identify open questions that warrant future exploration.

To begin, there's a notable inconsistency in the findings across the studies included in this thesis as well as earlier research. Specifically, my corpus analyses and corpus passage continuation experiment (chapters 3 and 4) found no influence of predictability on pronoun usage. However, the computational modeling and the meta-analysis (chapters 5 and 6) drew the contrary conclusion. A potential explanation for this discrepancy might be the nature of the contexts we explored. While the computational modeling and the meta-analysis examined a broader range of contexts, the corpus analyses and the continuation task narrowed their focus to three specific discourse relations: Narration, Contrast, and Result. The difference in subject re-mention biases between contexts of Narration and Contrast/Result, though statistically significant, did not exceed 25%, as shown in Figure 3.2 and Figure 4.2. Numerically, this difference in predictability induced by discourse relations is much smaller when compared to the difference observed with transfer-of-possession verbs and implicit causality verbs. For instance, Weatherford and Arnold (2021), who found an effect of predictability on pronoun usage, observed a substantial 57% distinction in subject re-mention rates between subject-biased and object-biased implicit causality contexts. Specifically, the subject was chosen as the next mention 76% of the time in subject-biased scenarios, whereas this figure dropped to a mere 19% when the object is the implicit cause. It is plausible that our relatively minor difference in referent predictability makes it hard to generate an effect on referring expression choice. Supporting this hypothesis, Demberg et al. (2023) recently demonstrated that the size of the effect on pronominalization strongly depends on the strength of referent predictability.

When predictability only slightly differs between conditions, the resulting effect on pronoun usage also hardly differs between conditions and, thus, requires a large sample for clear identification in studies, as suggested by our meta-analysis (Chapter 6).

An additional aspect of our corpus analyses and corpus passage continuation experiment that could potentially account for the absence of observed effects is their focus on written texts. An exploratory analysis in our meta-analysis (Appendix D.5) and recent investigations by Ye and Arnold (2023) and Demberg et al. (2023) indicate that the influence of predictability on pronoun usage is dependent on task modality. The influence of referent predictability tends to be more prominent in spoken and interactive communicative contexts. On the other hand, in standard written story continuation experiments in which there is not an obvious listener to tailor one's expressions to, participants simply complete sentences out of context. This lack of a clear conversational partner might not motivate them as much to behave as pragmatic speakers. It is therefore recommended to explore production in more naturalistic and interactive discourse settings, as online effects of audience design may not surface unless participants are presented with a genuine, engaging audience.

Overall, the findings of this thesis align more closely with the view that referent predictability does influence pronominalization choices, albeit in a modest way, as suggested by our computational modeling and meta-analysis. Speakers are rational and efficient, choosing more reduced forms like pronouns for more predictable referents. Nonetheless, it is clear that grammatical and structural factors, such as the grammatical role or topichood, are the main drivers of a speaker's choice of referential form. The influence of predictability, in comparison, is relatively minor, which might be explained by a trade-off in cognitive effort vs. communicative efficiency for speakers, as discussed at the end of our meta-analysis (Section 6.4). Such small effects can easily become eclipsed by other stronger factors.

Methodologically, each study in this thesis presents its own novelties. Our corpus analyses and corpus passage continuation are a step forward toward incorporating more naturalistic and realistic language into psycholinguistic research. By extracting passages from corpora, we gain the opportunity to investigate language production in a context that more closely mirrors how language is actually used by speakers and writers. Moreover, by using participants' self-annotations on referent choice and automatic labeling of referential form with Natural Language Processing techniques, we can examine a larger sample size without the tremendous post-experiment manual effort required for labeling. This approach is especially beneficial for larger studies that require more data and increased statistical power, such as those concerning referent predictability or other factors that may exhibit a small effect size. While stylistic features of corpus texts might pose challenges for examining very specific constructions, our computational model trained on masked coreference resolution can be applied to more diverse contexts and can be applied in a more scalable way. Concretely, we can employ a model like SpanBERT-coref or our own on new unlabeled texts across various genres to automatically detect mentions and identify coreference links between them, thus generating a larger dataset with annotations for analysis. This method might be particularly apt for investigations of languages like English and Mandarin Chinese, given their rich computational resources, in general and in particular in the realm of anaphoric reference and coreference resolution. Its application, however, is limited for many other languages, since we first require a sufficient amount of human-labeled data to train a model for satisfactory performance, ensuring that its predictions can serve as reliable standards and annotations. Fortunately, there have been efforts to explore various techniques to mitigate this limitation. These include transfer learning, which enables us to leverage knowledge gained from larger datasets (e.g., English datasets) to improve the processing of underrepresented languages, as well as data augmentation, which creates additional data by modifying existing data or incorporating data from diverse sources (e.g., Zoph et al., 2016; Fadaee et al., 2017). Future work should also consider alternative approaches that do not require annotation in the first place, akin to the use of language models for surprisal estimation (Pimentel et al., 2020). While the specific methods for achieving this are not yet fully defined, this avenue of research holds promise and merits further exploration. All in all, as Natural Language Processing continues to evolve swiftly, we remain optimistic about the prospects of future models enriching our understanding of linguistic phenomena.

In the last piece of this thesis, I draw inspiration from Galileo's pioneering gaze into the heavens. His blurry vision of Saturn through a primitive telescope symbolizes our endeavors to understand complex phenomena with the tools at our disposal. The journey of discovery, much like our use of pronouns, is punctuated by both clarity and ambiguity. Just as better telescopes revealed Saturn's true form, novel methods and resources, like those presented in this dissertation, will provide new insights into this linguistic question.

Bibliography

Alfaraz, G. (2015). Variation of overt and null subject pronouns in the Spanish of Santo Domingo. *Subject pronoun expression in Spanish: A cross-dialectal perspective*, pages 3–16.

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. Routledge, New York.

Ariel, M. (2001). Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, 8:29–87.

Armeni, K., Willems, R. M., and Frank, S. L. (2017). Probabilistic language models in cognitive neuroscience: Promises and pitfalls. *Neuroscience & Biobehavioral Reviews*, 83:579–588.

Arnold, J. E. (1998). *Reference form and discourse patterns*. PhD thesis, Stanford University.

Arnold, J. E. (2001). The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse processes*, 31(2):137–162.

Arnold, J. E. (2010). How speakers refer: The role of accessibility. *Language and Linguistics Compass*, 4(4):187–203.

Arnold, J. E., Brown-Schmidt, S., and Trueswell, J. (2007). Children's use of gender and order-of-mention during pronoun comprehension. *Language and cognitive processes*, 22(4):527–565.

Arnold, J. E. and Griffin, Z. M. (2007). The effect of additional characters on choice of referring expression: Everyone counts. *Journal of memory and language*, 56(4):521–536.

Arnold, J. E. and Tanenhaus, M. K. (2011). Disfluency effects in comprehension: How new information can become accessible. *The processing and acquisition of reference*, pages 197–217.

Asher, N., Asher, N. M., and Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Asher, N. and Vieu, L. (2005). Subordinating and coordinating discourse relations. *Lingua*, 115(4):591–610.

Aylett, M. and Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, 47(1):31–56.

Bader, M. and Portele, Y. (2019). The interpretation of German personal pronouns and d-pronouns. *Zeitschrift für Sprachwissenschaft*, 38(2):155–190.

Bagga, A. and Baldwin, B. (1998). Entity-based cross-document coreferencing using the vector space model. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

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

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*.

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons.

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

Brennan, S. E., Friedman, M. W., and Pollard, C. (1987). A centering approach to pronouns. In *25th Annual Meeting of the Association for Computational Linguistics*, pages 155–162.

Brocher, A., Chiriacescu, S. I., and von Heusinger, K. (2018). Effects of information status and uniqueness status on referent management in discourse comprehension and planning. *Discourse Processes*, 55(4):346–370.

Brown-Schmidt, S. and Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive science*, 32(4):643–684.

Bubic, A., Von Cramon, D. Y., and Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in human neuroscience*, page 25.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software*, 80:1–28.

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, 100(5):1–54.

Carlson, L., Marcu, D., and Okurovsky, M. E. (2001). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press.

Cheng, W. and Almor, A. (2019). A Bayesian approach to establishing coreference in second language discourse: Evidence from implicit causality and consequentiality verbs. *Bilingualism: Language and Cognition*, 22(3):456–475.

Chiriacescu, S. and von Heusinger, K. (2010). Discourse prominence and pe-marking in Romanian. *International Review of Pragmatics*, 2(2):298–332.

Choi, K.-Y. (2013). Hankwukeuy 3 inching cisi phyohyen ‘ku’ey kwuanhan soko [A look on 3rd person referring expression ku in Korean]. *Studies in Generative Grammar*, 23:527–558.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204.

Clark, K. and Manning, C. D. (2016). Improving coreference resolution by learning entity-level distributed representations. *arXiv preprint arXiv:1606.01323*.

Contemori, C. and Di Domenico, E. (2021). Microvariation in the division of labor between null-and overt-subject pronouns: the case of Italian and Spanish. *Applied Psycholinguistics*, 42(4):997–1028.

Crawley, R. A., Stevenson, R. J., and Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research*, 19(4):245–264.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Davis, F. and van Schijndel, M. (2021). Uncovering constraint-based behavior in neural models via targeted fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1159–1171, Online. Association for Computational Linguistics.

De Leeuw, J. R. (2015). jsPsych: A Javascript library for creating behavioral experiments in a Web browser. *Behavior research methods*, 47:1–12.

Demberg, V., Kravtchenko, E., and Loy, J. E. (2023). A systematic evaluation of factors affecting referring expression choice in passage completion tasks. *Journal of Memory and Language*, 130:104413.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Fedele, E. and Kaiser, E. (2015). Resolving null and overt pronouns in Italian: an experimental investigation of syntax–semantics interactions. *TLS: Proceedings of the 15th Texas Linguistic Society. Austin: University of Texas*, pages 53–72.

Ferretti, T. R., Rohde, H., Kehler, A., and Crutchley, M. (2009). Verb aspect, event structure, and coreferential processing. *Journal of memory and language*, 61(2):191–205.

Ferstl, E. C., Garnham, A., and Manouilidou, C. (2011). Implicit causality bias in English: A corpus of 300 verbs. *Behavior Research Methods*, 43(1):124–135.

Filiaci, F., Sorace, A., and Carreiras, M. (2014). Anaphoric biases of null and overt subjects in Italian and Spanish: a cross-linguistic comparison. *Language, Cognition and Neuroscience*, 29(7):825–843.

Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. (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)*, pages 878–883, Sofia, Bulgaria. Association for Computational Linguistics.

Frederiksen, A. T. and Mayberry, R. 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.

Fukumura, K. and Van Gompel, R. P. (2010). Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language*, 62(1):52–66.

Fukumura, K. and Van Gompel, R. P. (2011). The effect of animacy on the choice of referring expression. *Language and cognitive processes*, 26(10):1472–1504.

Fukumura, K. and van Gompel, R. P. (2015). Effects of order of mention and grammatical role on anaphor resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(2):501.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Gelormini-Lezama, C. (2018). Exploring the repeated name penalty and the overt pronoun penalty in Spanish. *Journal of Psycholinguistic Research*, 47:377–389.

Gelormini-Lezama, C. and Almor, A. (2011). Repeated names, overt pronouns, and null pronouns in Spanish. *Language and cognitive processes*, 26(3):437–454.

Gelormini-Lezama, C. and Almor, A. (2014). Singular and plural pronominal reference in Spanish. *Journal of Psycholinguistic Research*, 43:299–313.

Givón, T. (1983). *Topic continuity in discourse*. Amsterdam: John Benjamins.

Goikoetxea, E., Pascual, G., and Acha, J. (2008). Normative study of the implicit causality of 100 interpersonal verbs in Spanish. *Behavior Research Methods*, 40(3):760–772.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

Grüter, T., Rohde, H., and Schafer, A. J. (2017). Coreference and discourse coherence in L2: The roles of grammatical aspect and referential form. *Linguistic Approaches to Bilingualism*, 7(2):199–229.

Guan, S. and Arnold, J. E. (2021). The predictability of implicit causes: testing frequency and topicality explanations. *Discourse Processes*, pages 1–27.

Gundel, J. K. (1988). Universals of topic-comment structure. *Studies in syntactic typology*, 17(1):209–239.

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

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Harari, M. B., Parola, H. R., Hartwell, C. J., and Riegelman, A. (2020). Literature searches in systematic reviews and meta-analyses: A review, evaluation, and recommendations. *Journal of Vocational Behavior*, 118:103377.

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., and Welch, V. A. (2019). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

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

Holler, A. and Suckow, K., editors (2016). *How clausal linking affects noun phrase salience in pronoun resolution*, pages 61–86. De Gruyter, Berlin, Boston.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

Hwang, H. (2023a). Choice of nominative and topic markers in Korean discourse. *Quarterly Journal of Experimental Psychology*, 76(4):905–921.

Hwang, H. (2023b). The influence of discourse continuity on referential form choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 49(4).

Hwang, H. and Lam, S. Y. (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, H., Lam, S. Y., Ni, W., and Ren, H. (2022). The role of grammatical role and thematic role predictability in reference form production in Mandarin Chinese. *Frontiers in psychology*, 13.

Jaeger, T. and Levy, R. (2006). Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19.

Jaeger, T. F. and Buz, E. (2017). Signal reduction and linguistic encoding. *The Handbook of Psycholinguistics*, pages 38–81.

Jäger, L. A., Engelmann, F., and Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, 94:316–339.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Joshi, M., Levy, O., Zettlemoyer, L., and Weld, D. (2019). BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. *Typological studies in language*, 45:229–254.

Kaiser, E. (2010). Investigating the consequences of focus on the production and comprehension of referring expressions. *International Review of Pragmatics*, 2(2):266–297.

Kehler, A. (2002). *Coherence, reference, and the theory of grammar*. CSLI publications Stanford, CA.

Kehler, A. (2019). Coherence Relations. In *The Oxford Handbook of Event Structure*. Oxford University Press.

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

Kehler, A. and Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39(1-2):1–37.

Kehler, A. and Rohde, H. (2019). Prominence and coherence in a Bayesian theory of pronoun interpretation. *Journal of Pragmatics*, 154:63–78.

Keysar, B., Barr, D. J., Balin, J. A., and Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11(1):32–38.

Kim, H. (1990). Continuity of action and topic in discourse. *Japanese/Korean linguistics*, pages 79–96.

Konuk, G. and von Heusinger, K. (2021). Discourse prominence in Turkish: The interaction of grammatical function and semantic role. In Khomchenkova, I., Sinitsyna, Y., and Tatevosov, S., editors, *Proceedings of the 15th Workshop on Altaic Formal Linguistics (WAFL 15)*, MIT working papers in linguistics, pages 109–120. Dept., Cambridge, MA.

Kravtchenko, E. (2022). *Integrating pragmatic reasoning in an efficiency-based theory of utterance choice*. PhD thesis, Universität des Saarlandes.

Kuperberg, G. R. and Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, 31(1):32–59.

Lam, S.-Y. and Hwang, H. (2022). How does topicality affect the choice of referential form? Evidence from Mandarin. *Cognitive Science*, 46(10):e13190.

Lambrecht, K. (1996). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*, volume 71. Cambridge University Press.

Lau, S. H. and Hwang, H. (2016). The effects of frequency on pronoun production. *Journal of Cognitive Science*, 17(4):547–569.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Lee, K., He, L., and Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Leventhal, G. (1973). Effect of sentence context on word perception. *Journal of Experimental Psychology*, 101(2):318.

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

Levy, R. and Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19:849.

Liao, X. (2022). Coherence-driven predictability and referential form: Evidence from English corpus data. In Gutzmann, D. and Repp, S., editors, *Proceedings of Sinn und Bedeutung 26*, pages 544–556, Universität zu Köln.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and speech*, 6(3):172–187.

Lindemann, S.-I., Mada, S., Sasu, L., and Matei, M. (2020). Thematic role and grammatical function affect pronoun production. *ExLing 2020*, page 113.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Mahowald, K., James, A., Futrell, R., and Gibson, E. (2016). A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*, 91:5–27.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Mayol, L. (2018). Asymmetries between interpretation and production in Catalan pronouns. *Dialogue & Discourse*, 9(2):1–34.

McCoy, K. E. and Strube, M. (1999). Generating anaphoric expressions: Pronoun or definite description? In *Proceedings of ACL workshop on Discourse and Reference Structure*, University of Maryland.

McDonald, J. L. and MacWhinney, B. (1995). The time course of anaphor resolution: Effects of implicit verb causality and gender. *Journal of Memory and Language*, 34(4):543–566.

Medina Fetterman, A. M., Vazquez, N. N., and Arnold, J. E. (2022). The effects of semantic role predictability on the production of overt pronouns in Spanish. *Journal of Psycholinguistic Research*, pages 1–26.

Modi, A., Anikina, T., Ostermann, S., and Pinkal, M. (2016). InScript: Narrative texts annotated with script information. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3485–3493, Portorož, Slovenia. European Language Resources Association (ELRA).

Modi, A., Titov, I., Demberg, V., Sayeed, A., and Pinkal, M. (2017). Modeling semantic expectation: Using script knowledge for referent prediction. *Transactions of the Association for Computational Linguistics*, 5:31–44.

Orita, N., Vornov, E., Feldman, N., and Daumé III, H. (2015). Why discourse affects speakers' choice of referring expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1639–1649.

Özge, U., Özge, D., and Von Heusinger, K. (2016). Strong indenites in Turkish, referential persistence, and salience structure. *Empirical perspectives on anaphora resolution*, pages 169–191.

Patterson, C., Schumacher, P. B., Nicenboim, B., Hagen, J., and Kehler, A. (2022). A Bayesian approach to German personal and demonstrative pronouns. *Frontiers in psychology*, 12:6296.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Piantadosi, S. T., Tily, H., and Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Pimentel, T., Maudslay, R. H., Blasi, D., and Cotterell, R. (2020). Speakers fill lexical semantic gaps with context. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4004–4015, Online. Association for Computational Linguistics.

Poesio, M., Artstein, R., Uryupina, O., Rodriguez, K., Delogu, F., Bristot, A., and Hitzeman, J. (2013). The ARRAU Corpus of Anaphoric Information. *Linguistic Data Consortium*.

Portele, Y. and Bader, M. (2020). Coherence and the interpretation of personal and demonstrative pronouns in German. In *Information Structuring in Discourse*, pages 24–55. Brill.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rohde, H. (2008). *Coherence-driven effects in sentence and discourse processing*. PhD thesis, University of California, San Diego.

Rohde, H. and Kehler, A. (2014). Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, 29(8):912–927.

Rosa, E. C. (2015). *Semantic role predictability affects referential form*. PhD thesis, The University of North Carolina at Chapel Hill.

Rosa, E. C. and Arnold, J. E. (2017). Predictability affects production: Thematic roles can affect reference form selection. *Journal of Memory and Language*, 94:43–60.

Schutze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Solstad, T. and Bott, O. (2022). On the nature of implicit causality and consequentiality: the case of psychological verbs. *Language, Cognition and Neuroscience*, pages 1–30.

Song, J. and Kaiser, E. (2020). Forward-looking effects in subject pronoun interpretation: What comes next matters. In *CogSci*.

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, 9(8):311–327.

Stevenson, R. J., Crawley, R. A., and Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, 9(4):519–548.

Tily, H. and Piantadosi, S. (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*.

Upadhye, S., Bergen, L., and Kehler, A. (2020). Predicting reference: What do language models learn about discourse models? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 977–982.

van Deemter, K. (1990). Forward references in natural language. *Journal of Semantics*, 7(3):281–300.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Verhagen, V., Mos, M., Backus, A., and Schilperoord, J. (2018). Predictive language processing revealing usage-based variation. *Language and cognition*, 10(2):329–373.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Vogels, J. (2019). Both thematic role and next-mention biases affect pronoun use in Dutch. In *CogSci*, pages 3029–3035.

Weatherford, K. C. and Arnold, J. E. (2021). Semantic predictability of implicit causality can affect referential form choice. *Cognition*, 214:104759.

Webber, B., Prasad, R., Lee, A., and Joshi, A. (2019). The Penn Discourse Treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., et al. (2013). Ontonotes release 5.0. *Linguistic Data Consortium*.

Wurm, L. H. and Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of memory and language*, 72:37–48.

Ye, Y. and Arnold, J. (2023). Implicit causality affects pronoun use for speakers (but not writers). Poster presented at 36th Annual Conference on Human Sentence Processing, University of Pittsburgh.

Zarcone, A., Van Schijndel, M., Vogels, J., and Demberg, V. (2016). Salience and attention in surprisal-based accounts of language processing. *Frontiers in psychology*, 7:844.

Zerkle, S. A. and Arnold, J. E. (2019). Does pre-planning explain why predictability affects reference production? *Dialogue & Discourse*, 10(2):34–55.

Zerkle, S. A., Rosa, E. C., and Arnold, J. E. (2017). Thematic role predictability and planning affect word duration. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 8(1).

Zhan, M., Levy, R., and Kehler, A. (2020). Pronoun interpretation in Mandarin Chinese follows principles of Bayesian inference. *Plos one*, 15(8):e0237012.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Appendix A

MORE INFORMATION ABOUT 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 reproduce 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

First, we examine how often different thematic roles are mentioned next in transfer-of-possession scenarios. Secondly, we study the use of pronouns in these scenarios while controlling for grammatical roles. This aims to answer 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.1.1. 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 A.1.

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 A.1: 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 identified using the annotation of predicate-argument structure (see Table 3.3 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 A.1, where mentions marked with the same color refer to the same entity, and they are in the same reference chain and annotated with the same reference ID.

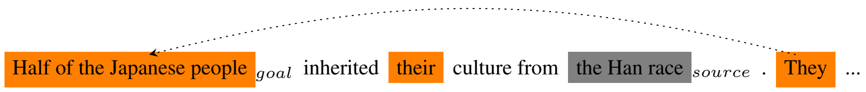

Half of the Japanese people$_{goal}$ inherited their culture from the Han race$_{source}$. They ...

Figure A.1: A context which continues with the goal referent.

A.1.2. 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 A.2 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 A.2).

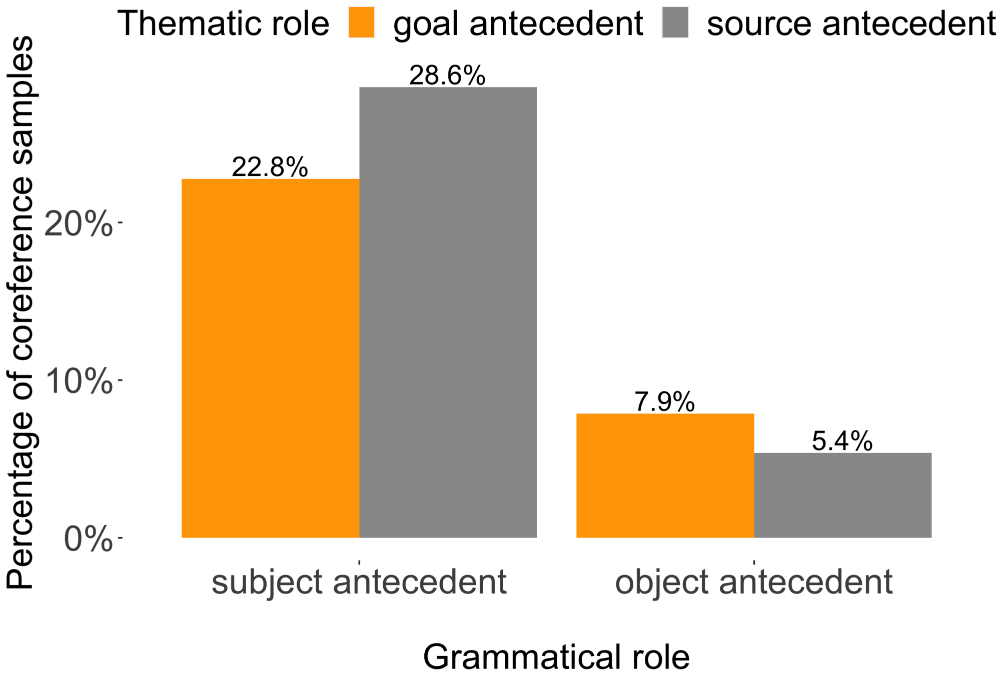


Figure A.2: Percentage of goal and source continuations in two transfer-of-possession contexts.

As we can see in Figure A.2, for subject antecedents, sources were mentioned more frequently than goals, while the opposite was found for object antecedents. Overall, the chi-squared test show 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: $X^2(1) = .71$,

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 A.2: Number of coreference samples automatically retrieved for transfer-of-possession verbs.

p = .4).

Figure A.3 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 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.

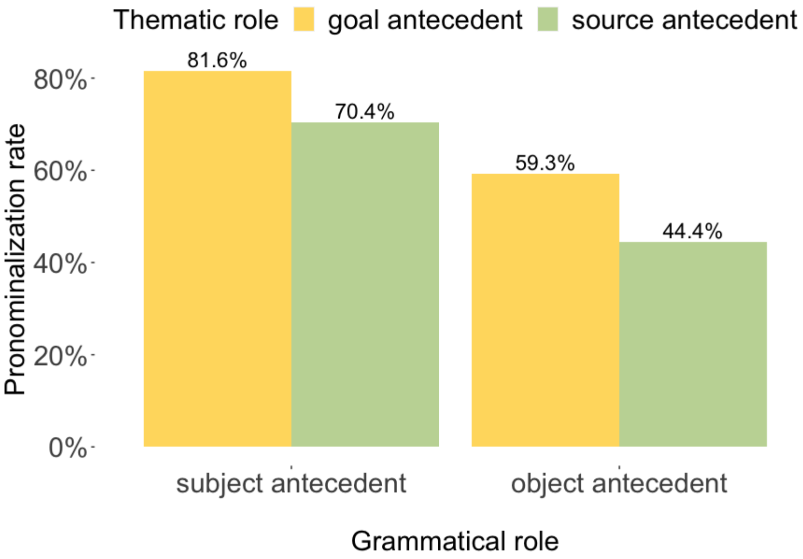


Figure A.3: Pronominalization rate in each category for transfer-of-possession verbs.

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[1].

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., 30, rather than concrete ones as in the items of psycholinguistic experiments, e.g. 7. When used in other senses, transfer-of-possession verbs do not necessarily elicit a focus on the goal.[2]

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

A.2. Analysis 2: implicit causality verbs

This section presents our analyses 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., Stimulus-Experiencer verbs; *scare*, *surprise*) more likely to continue with the subject, and object-biased verbs (e.g., Experiencer-Stimulus verbs; *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.2.1. Method

These analyses followed a similar methodology as those 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

[1] 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.

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

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).[3] All implicit causality verbs used in this extraction are listed in Appendix A.3. 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 reference chain of the next mention was compared against that of the two antecedents.

A.2.2. 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 previously for the analysis with transfer-of-possession verbs (Section A.1.2). Figure A.4 shows the percentage of references to the subject and object antecedents (the raw numbers are presented in Table A.3).

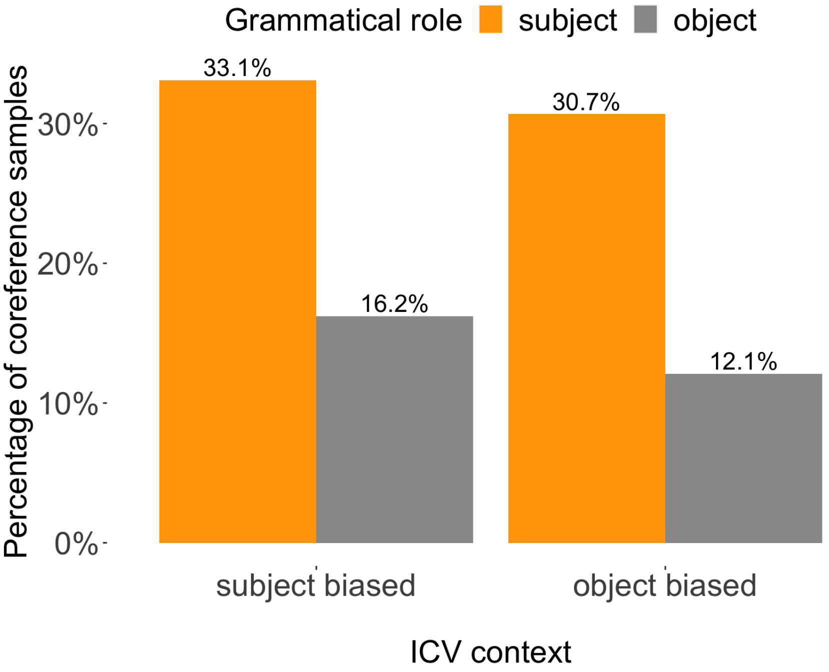


Figure A.4: Percentage of continuations in subject-biased and object-biased implicit causality verb contexts.

[3] 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).

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

Table A.3: 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 reproduce the contrasting likelihoods of next mention.

We present pronominalization rates in Figure A.5 for extra information. In Figure A.5, 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 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 31. 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.

(31) Russian Foreign Minister Igor Ivanov congratulated Kostunica on his election

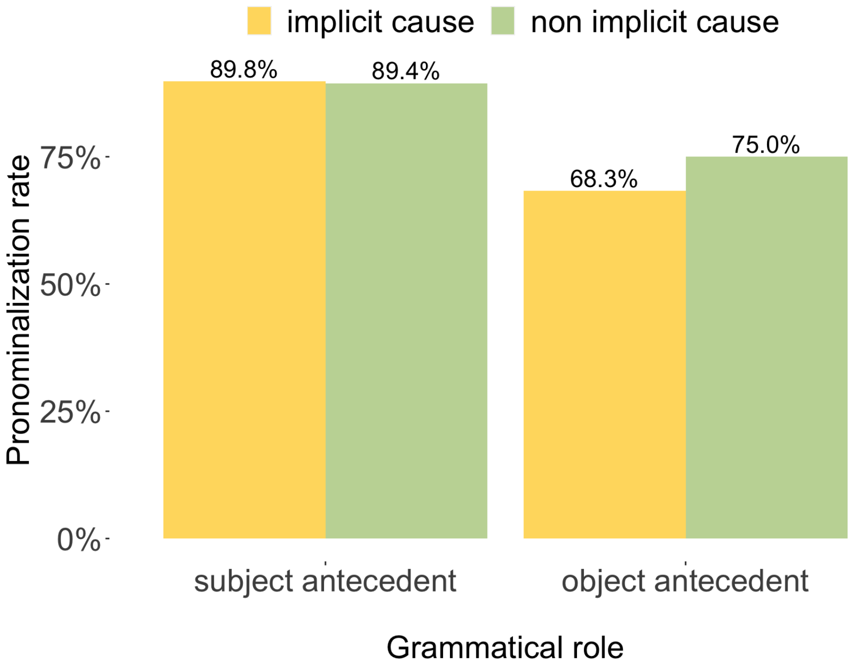


Figure A.5: Pronominalization rate in each category for subject-biased and object-biased implicit causality verb contexts.

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 end-point such as the location in 32, which reduces the probability of continuing with goal referents. Implicit causality verbs have an analogous problem (see the object *the broader selection* in 33). This makes it difficult to control for the topicality of arguments.

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

(33) 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.3. 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

Appendix B

MORE INFORMATION ABOUT THE ANALYSES WITH DISCOURSE RELATIONS

B.1. 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 B.1: 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.2. 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 B.1 for examples).

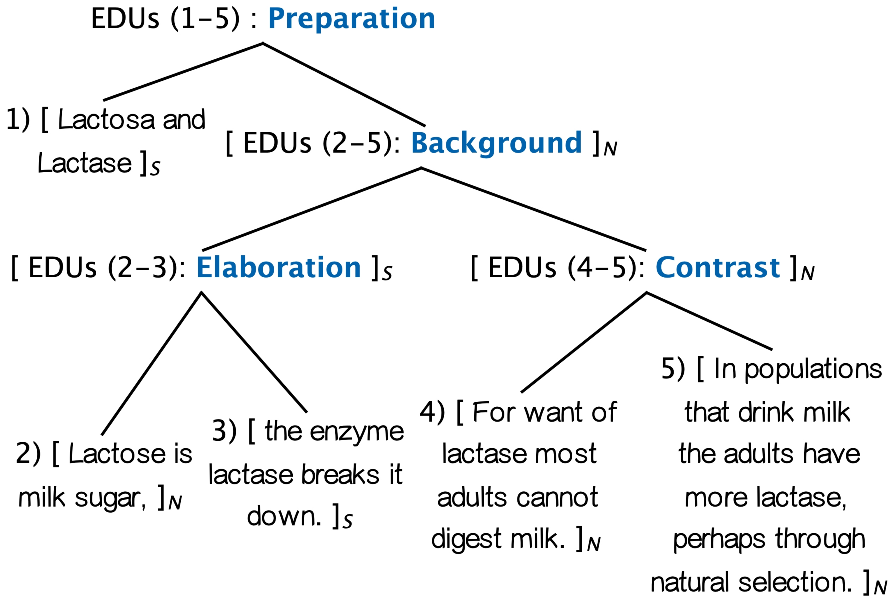

EDUs (1–5) : Preparation
1) [Lactosa and Lactase]$_S$
[EDUs (2–5): Background]$_N$
[EDUs (2–3): Elaboration]$_S$
[EDUs (4–5): Contrast]$_N$
2) [Lactose is milk sugar,]$_N$
3) [the enzyme lactase breaks it down.]$_S$
4) [For want of lactase most adults cannot digest milk.]$_N$
5) [In populations that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

Figure B.1: Graphical representation of an RST analysis, with nucleus/satellite annotated.

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 rhetorical relations. For instance, while example 34 is annotated as a “Result” relation, example 35 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 obtained 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 Occasion 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.

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

(35) 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.]$_N$

B.3. 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 B.2: 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 B.3: RST-DT: Counts of samples in each coreference type by discourse relation.

B.4. 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.[1] This factor could potentially

[1] Given the limited number of samples with pronominal antecedents in RST-DT (see Table B.4), we did not conduct further analysis to explore their potential influences on our results.

lead to varying levels of topicality among the referents across different relations.[2] 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.

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

Table B.4: Raw pronoun data for coreference samples in RST-DT. PRO stands for pronoun. NM is abbreviation for next-mention.

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 B.5, 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.

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

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

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 B.5: Pronominalization of subject re-mentions in OntoNotes: mixed-effects logistic regression model with the next mention being a pronoun as the dependent measure.

Appendix C

MORE INFORMATION ABOUT THE COMPUTATIONAL MODELING

C.1. Method: details

For simplicity, both in training and evaluation, we never mask mentions which are embedded in another mention (e.g., “the bride” in “the mother of the bride”), since that would cover information relevant to the larger mention. In case we mask a mention that includes another mention, we discard the latter from the set of mentions for which to compute a prediction.

For evaluation on development data, to find the best models across training epochs and hyperparameters, we use a quicker but more coarse-grained method than that used for evaluation on test data to assess performances on masked mentions. We mask a random sample (10%; independently of the percentage used during training) of mentions in each document, compute evaluation scores and get the average of these across 5 iterations (i.e., with different samples of mentions masked). Although in this setup masks could potentially interfere with each other, and we will not have masked predictions for all mentions, overall this method will give us a good enough representation of the model’s performances on masked mentions, while being quick to compute.

When evaluating antecedent prediction, we skip the first mention in a document as this is a trivial prediction (no antecedent).

C.2. Evaluation

C.2.1. Complete results on OntoNotes

Table C.1 reports MUC, B^3 and CEAF scores (precision, recall and F1), for the $\mathbf{M}_{\mathrm{u}}$ and $\mathbf{M}_{\mathrm{m}}$. The results are overall comparable between the two systems across all metrics.

model	mentions	MUC			B^3			CEAF			Average of metrics		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
$\mathbf{M}_{\mathrm{u}}$	predicted	.84	.83	.84	.76	.75	.76	.75	.71	.73	.78	.77	.77
	gold	.95	.91	.93	.87	.86	.86	.92	.77	.84	.91	.85	.88
$\mathbf{M}_{\mathrm{m}}$	predicted	.84	.83	.83	.76	.75	.75	.74	.71	.73	.78	.76	.77
	gold	.95	.93	.94	.86	.88	.87	.92	.78	.85	.91	.86	.88

Table C.1: Results on OntoNotes test data (English) in document-level coreference resolution (only with unmasked mentions); P, R, F1 = precision, recall, F1 scores.

Figures C.1 and C.2 report the antecedent prediction results, using gold mention boundaries, of the $\mathbf{M}_{\mathrm{m}}$ and $\mathbf{M}_{\mathrm{u}}$ considering fine-grained distinctions across mention types than what reported in Figure 5.6 of the paper. Concretely, we divide pronouns, into first-, second- and third-person pronouns, as well as treating demonstratives (e.g., “that”) as a separate category (DEM).

We subdivide pronouns in this way because they are quite heterogeneous: first- and second-person pronouns are comparatively rigid (typically referring to the speaker and addressee), and are used oftentimes within a quotation (e.g. Asked why senators were giving up so much, New Mexico Sen. Pete Dominici, [...] said, “[**We**]’re looking like idiots [...]”); and demonstrative pronouns tend to be more difficult cases in OntoNotes, for instance referring to the head of verbal phrases (e.g. [...] their material life will no doubt be a lot less taken of when compared to the usual both parents or one parent at home situation. [**This**] is indisputable). Overall, for masked mentions, precision is high across pronouns, and highest among pronoun types for third-person pronouns. For unmasked mentions, the hardest cases are demonstrative pronouns.

We also report these results looking at predictions with predicted (i.e., identified by the system) mention boundaries. These are displayed in Figure C.4 and C.3 for $\mathbf{M}_{\mathrm{u}}$ and $\mathbf{M}_{\mathrm{m}}$, respectively. While results are generally better with gold mention boundaries, the trends stay the same across the two setups for both masked and unmasked mentions.

Finally, in Figure C.5 we report the results looking at a variant of $\mathbf{M}_{\mathrm{m}}$ where instead of substituting mentions with one `[MASK]` token we use a sequence of

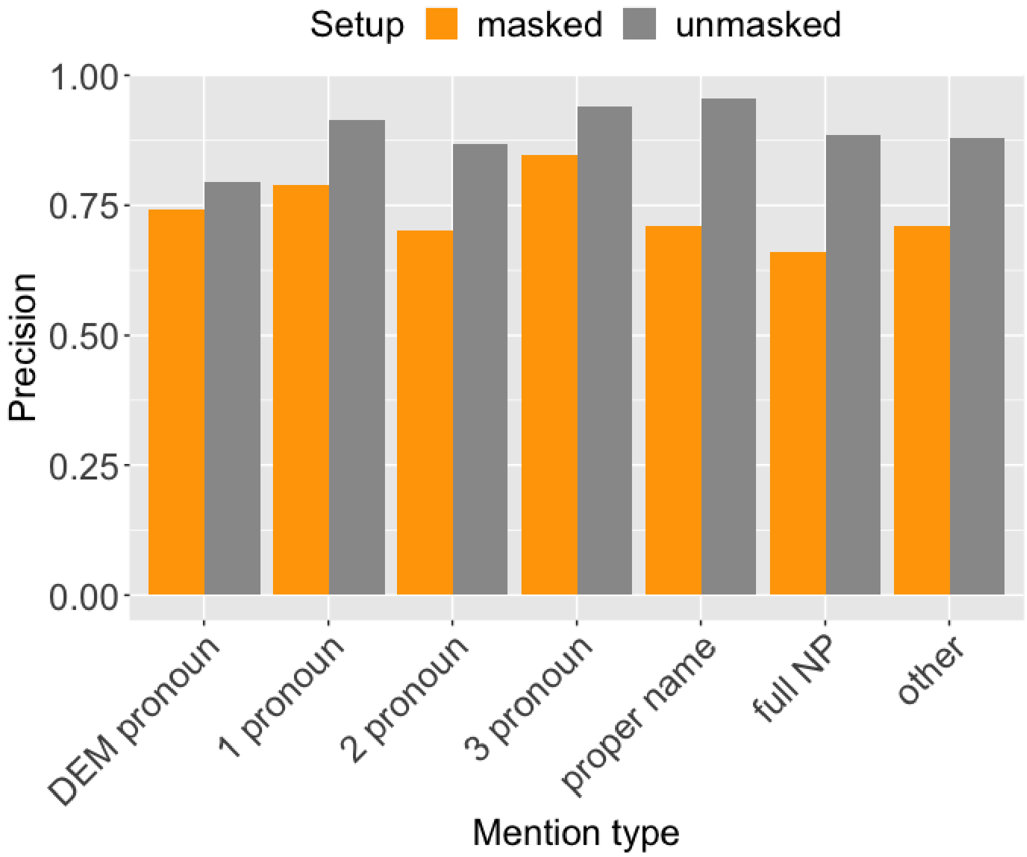


Figure C.1: Antecedent precision for $\mathbf{M}_{\mathrm{m}}$ across more fine-grained mention types, for masked and unmasked mentions.

three. This is to verify whether the use of a single token biases the system to be better on one-token mentions. The results show that this is not the case, as the trends found with the one-token masking are the same as those with the three-tokens masking: In particular, when a third-person pronoun is used the antecedent is still easier to predict than when a proper name is used, and even less than a full NP.

C.2.2. Comparison to human predictions

To elicit human judgments of referent predictability, Modi et al. (2017) relied on mention heads rather than the complete mention (e.g., “supermarket” in “the supermarket”). For one, they constructed the cloze task by cutting a text right before the head of the target mention (e.g., before “supermarket”), thus leaving part of the mention visible (e.g., “the” in this case). Moreover, they indicated

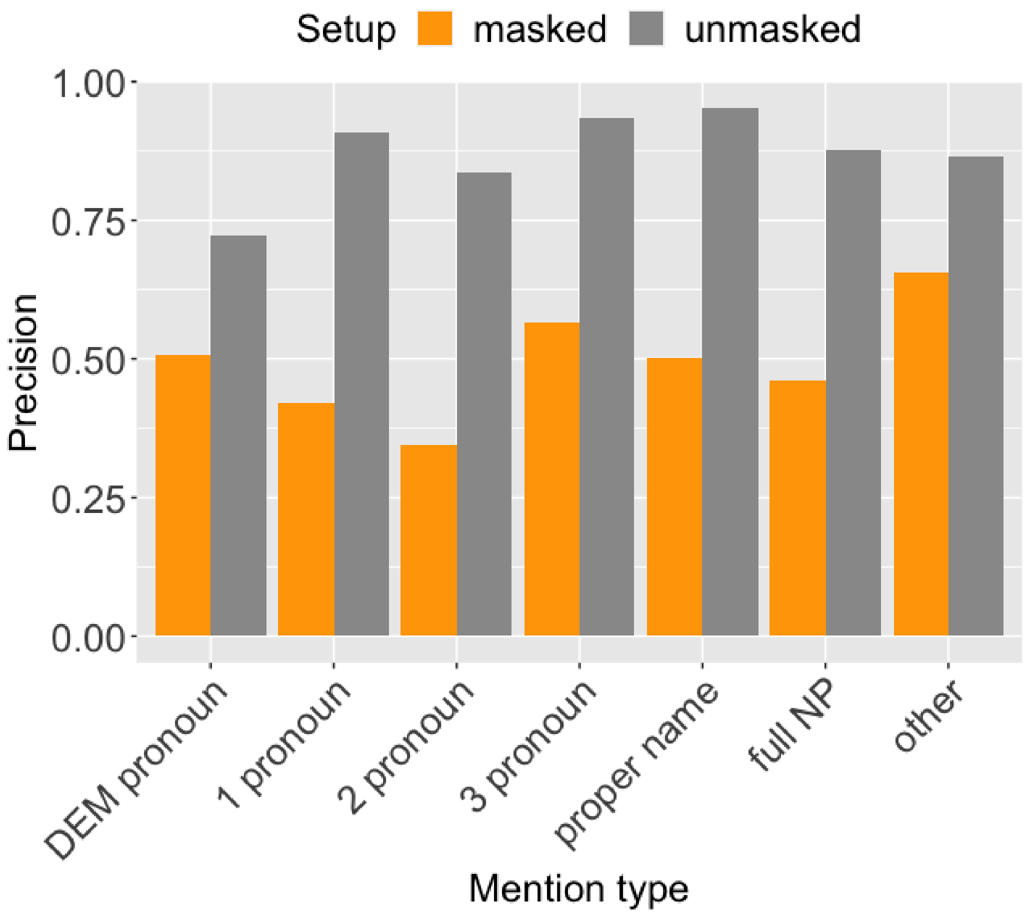


Figure C.2: Antecedent precision scores with gold mentions of the model $\mathbf{M}_\mathrm{u}$ across different mention types, for both masked and unmasked mentions.

candidate antecedent mentions for the human participants to consider, by listing again only the mention heads.

To make this task suitable for standard coreference resolution we need to identify the full mention boundaries belonging to each head (not given in the original annotations). To that end we rely on 'noun chunks' identified by the spaCy library, amended by a number of heuristics, for an estimated 91% accuracy (estimated by manually checking a sample of 200 mentions for correctness). We use the identified mention boundaries as gold mention boundaries exactly as in our OntoNotes setup (Section 5.2). However, different from our OntoNotes setup, we mask only the head of the target mention, exactly as in the human cloze task.

Table C.2 reports the results of $\mathbf{M}_\mathrm{m}$ on the data by Modi et al. (2017). We deploy the system in two setups: 1) Using just the left context of the target mention,

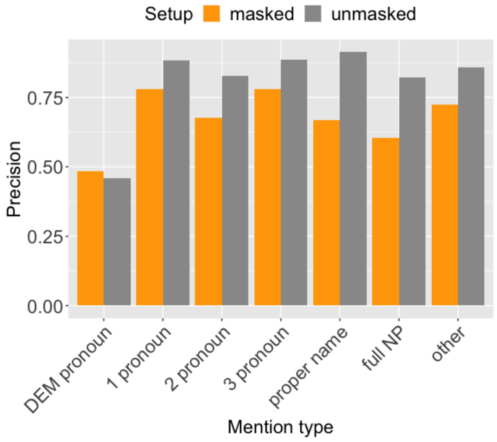


Figure C.3: Antecedent precision scores with predicted mentions of the model $\mathbf{M}_{\mathrm{m}}$ across different mention types, for both masked and unmasked mentions.

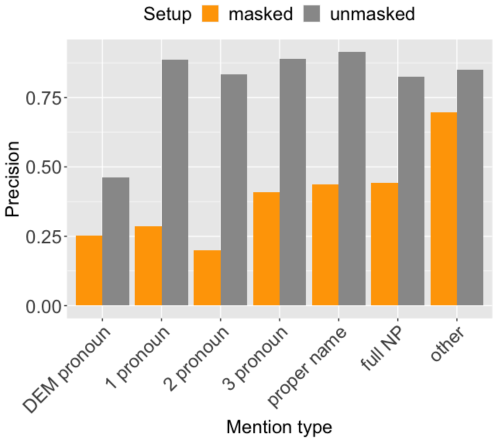


Figure C.4: Antecedent precision scores with predicted mentions of the model $\mathbf{M}_{\mathrm{u}}$ across different mention types, for both masked and unmasked mentions.

mimicking the setup used to elicit the human judgments, and 2) Using both the left and right context of the mention. In both cases, our results improve over those reported by Modi et al. (2017) for their best model, indicating that through our method we obtain better proxies for human discourse expectations.

$\mathbf{M}_{\mathrm{m}}$'s predictions are more aligned to those of humans when accessing both sides of the context than with only the left context, in spite of the second setup more closely resembling that used for the human data collection. Since information in the following context could not influence the human judgements (it was not available), we take this result to indicate that $\mathbf{M}_{\mathrm{m}}$ works generally better when deployed in a setup that is closer to that used during its training (recall that in training it never observed texts cropped after a masked mention), leading to suboptimal predictions when only the left context is used. We plan to explore this further in future work, by experimenting with variants to the training setup or different architectures (e.g., auto-regressive) that may improve the model's ability to resolve mentions based only on their previous contexts.

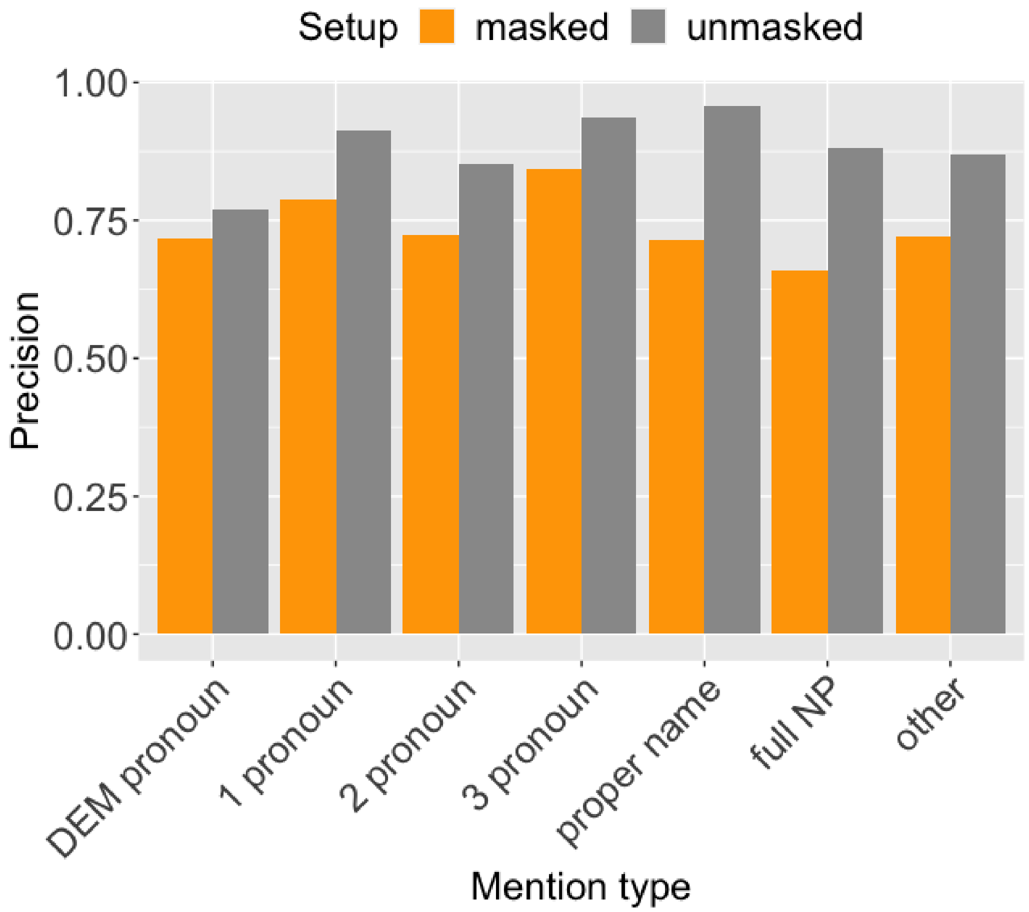


Figure C.5: Antecedent precision comparison of masked cloze task across mention types, for both masked and unmasked mentions, with three [MASK] tokens.

C.3. Predictability and mention form

C.3.1. Regression with both surprisal and entropy

In addition to surprisal, Tily and Piantadosi (2009) and Modi et al. (2017) also consider the uncertainty over competitors as a feature that captures some aspect of predictability. This uncertainty, more precisely **entropy**, is defined as *expected* surprisal:

$$\text{entropy}(x) := \sum_{e \in E_x} P(E_x = e \mid c_x) \cdot \text{surprisal}(x)$$

Entropy will be low if the probability mass centers on one or a few entities, and high if the probability is more evenly distributed over many entities, regardless of

	accuracy	relative accuracy w.r.t human top guess	JSD
$\mathbf{M}_{\mathrm{m}}$ left only	.54	.50	.46
$\mathbf{M}_{\mathrm{m}}$ left + right	.74	.64	.39
Modi et al. (2017)	.62	.53	.50

Table C.2: Evaluation of $\mathbf{M}_{\mathrm{m}}$ against human guesses using different amounts of context, in terms of average relative accuracy with respect to human top guess, as well as average Jensen-Shannon divergence (smaller is better) between the probability distribution of human predictions and model predictions.

which entity is the correct one.

In principle, entropy and surprisal capture genuinely different aspects of predictability; for instance, when the model is confidently incorrect, surprisal is high while entropy is low. However, in our data, entropy and surprisal are highly correlated (r_s = .87, $p < .001$). We did not fit regression models with both by residualising entropy to eliminate the collinearity, as our precedents did, because of the shortcomings of treating residualisation as remedy for collinearity (Wurm and Fisicaro, 2014). Instead, we define predictability primarily by surprisal (Uniform Information Density, Levy and Jaeger 2007) in our main analysis, and report the regression with both surprisal and the non-residualised entropy as a supplementary analysis. Note that we do not intend to interpret the coefficient of surprisal or entropy in this analysis (this is not possible because they are collinear), but rather to test whether surprisal and entropy still improve goodness-of-fit to the data on top of many other shallow linguistic features. Again, the shallow features themselves may capture aspects of entropy, and are indeed correlated with entropy (though all $r < .50$).

Table C.3 shows that both surprisal and entropy still matter for mention choice when controlling for the other factors, even though their statistical significance might be undermined due to the collinearity between them. Compared to the model with predictability primarily formulated as surprisal, similar effect patterns are found with entropy added, except that entropy seems to be better at distinguishing between pronoun vs. non-pronouns, and as the contexts become more uncertain, proper names and full NPs are roughly equally favored ($z = -.88$, $p = .38$) over pronouns after controlling for other variables.

C.3.2. More analyses results

Figure C.6 displays the predictions of mention type from the multinomial regression model, based on shallow features as well as surprisal. Each point represents a division of probability between the three levels of mention type. The

		Proper name				Full NP						
		β	s.e.	z	p	β	s.e.	z	p	LRχ2	df	p_χ2
Intercept		-.61	.03	-22.8	-	-.25	.02	-10.3	-			
surprisal		.16	.03	5.1	*	.33	.03	12.4	*	330.3	2	*
entropy		.42	.03	14.3	*	.45	.03.	16.8	*	349.4	2	*
Intercept		-.29	.07	-4.2	-	-.01	.07	-0.2	-			
distance		3.01	.12	24.4	*	3.00	.12	23.1	*	1294.3	2	*
frequency		.09	.03	3.3	*	-.12	.03	-3.6	*	37.1	2	*
antecedent	previous subject	-1.29	.09	-13.7	*	-1.06	.08	-13.4	*	346.6	2	*
mention	subject	.12	.07	1.7	.1	-0.44	.07	-6.8	*	72.9	2	*
antecedent type	proper name	1.79	.08	22.8	*	.42	.09	4.7	*	1730.7	2	*
	full NP	-.16	.08	-2.0	*	1.20	.07	18.3	*			
surprisal		-.01	.04	-0.2	.8	.17	.03	6.0	*	52.5	2	*
entropy		.23	.03	6.7	*	.25	.03	8.4	*	77.5	2	*

Table C.3: Two Multinomial logit models predicting mention type (baseline level is “pronoun”), based on 1) surprisal & entropy and 2) shallow linguistic features + surprisal & entropy. * marks predictors that are significant at the .05 alpha level.

corners of the triangle correspond to probability 1 for one outcome level and 0 for the other two, and the centre corresponds to probability 1/3 for each. Our model clusters most of the true pronouns (red) in the bottom left, and true full NPs (blue) in the bottom right, true proper names (green) at the top. Besides, many datapoints obtain similar division of probability, suggesting that some of them share similar pattern of features (recency, frequency etc.).

Table C.4 shows two linear regression models predicting mention length quantified in terms of number of tokens for each non-pronominal mention (proper name, full NP).

In the first model, we regress mention length (number of tokens) on surprisal alone. In the second one, the coefficient of surprisal decreases a bit with other shallow linguistic features added. F-tests are carried out to test if each predictor improves the fits to the data. In the fuller model, ”frequency” and ”antecedent type” are tested to significantly improve the model fit, above and over which surprisal still matters for mention type: longer non-pronominal expressions are favoured with surprisal increasing. We show similar effect pattern with two linear regression models predicting mention length alternatively measured in terms of characters, in Table C.6.

Table C.5 displays two linear regression models predicting mention length measured in terms of number of characters for each masked mention (including pronominal mentions). Compared to models for non-pronominal mentions, features like ”distance”, ”antecedent type” matter more when predicting the mention length with pronouns included, suggesting that these features better identify the distinction between pronouns vs. non-pronouns, but probably not between shorter

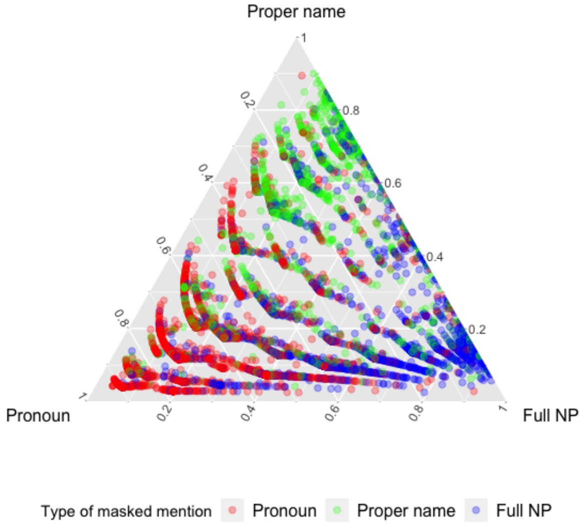


Figure C.6: Ternary probability plot. Each point represents predicted probabilities from the multinomial logit model with shallow features and surprisal predicting the three levels of mention type, which sum to 1.

and longer non-pronominal expressions.

Table C.7 and C.8 add results from likelihood-ratio chi-square tests and F-tests to Table 5.3 in the main text. All variables are tested to significantly improve goodness-of-fit to the data, except the feature ”target mention is subject” in predicting mention length (number of tokens).

		β	s.e.	t	p_t	F	df	p_F
Intercept		2.53	.04	64.24	-			
surprisal		.18	.04	5.10	*	26.00	1	*
Intercept		2.69	.09	30.72	-			
distance		-.02	.03	-0.71	.48	.51	1	.48
frequency		-.31	.05	-6.87	*	47.17	1	*
antecedent	previous subject	-.14	.15	-.93	.35	.87	1	.35
mention	subject	.10	.09	1.05	.29	1.11	1	.29
antecedent type	proper name	-.96	.11	-8.71	*	77.26	2	*
	full NP	.16	.10	1.56	.12			
surprisal		.16	.04	4.44	*	19.74	1	*

Table C.4: Two Linear regression models predicting mention length for each masked non-pronominal mention, based on 1) surprisal alone and 2) shallow linguistic features + surprisal.

		β	s.e.	t	p_t	F	df	p_F
Intercept		8.30	.11	75.68	-			
surprisal		1.40	.11	12.76	*	162.83	1	*
Intercept		8.01	.21	37.68	-			
distance		1.00	.11	8.87	*	78.64	1	*
frequency		-.79	.11	-6.93	*	48.04	1	*
antecedent	previous subject	-3.02	.29	-10.58	*	111.88	1	*
mention	subject	-.02	.25	-.06	.95	.004	1	.95
antecedent type	proper name	-.40	.30	-1.32	.19	48.45	2	*
	full NP	2.05	.26	7.86	*			
surprisal		.95	.11	8.57	*	73.52	1	*

Table C.5: Two Linear regression models predicting the character count (without space) for each masked mention, based on 1) surprisal alone and 2) shallow linguistic features + surprisal. F-test compares the fits of models.

		β	s.e.	t	p_t	F	df	p_F
Intercept		12.19	.18	67.95	-			
surprisal		.88	.16	5.54	*	30.69	1	*
Intercept		12.93	.40	32.47	-			
distance		-.08	.15	-.56	.58	.31	1	.58
frequency		-1.68	.20	-8.18	*	66.83	1	*
antecedent	previous subject	-.74	.67	-1.11	.27	1.24	1	.27
mention	subject	.63	.42	1.49	.14	2.21	1	.14
antecedent type	proper name	-4.16	.50	-8.26	*	63.40	2	*
	full NP	.41	.47	.89	.37			
surprisal		.76	.16	4.75	*	22.58	1	*

Table C.6: Two Linear regression models predicting the character count (without space) for each masked non-pronominal mention, based on 1) surprisal alone and 2) shallow linguistic features + surprisal. F-test compares the fits of models.

		Proper name			Full NP					
		β	s.e.	z	β	s.e.	z	LRχ^2	df	p_{χ^2}
Intercept		-.63	.03	-23.77	-.26	.02	-10.93			
surprisal		.31	.03	9.56	.47	.03	16.44	330.28	2	*
Intercept		-.24	.07	-3.60	.04	.07	0.58			
distance		3.13	.12	25.40	3.10	.12	25.23	1466.81	2	*
frequency		.09	.03	3.11	-.13	.03	-3.78	37.35	2	*
antecedent	previous subject	-1.31	.09	-13.91	-1.10	.08	-13.70	362.02	2	*
mention	subject	.07	.07	1.00	-0.50	.06	-7.71	82.57	2	*
antecedent type	proper name	1.78	.08	22.83	.41	.09	4.62	1723.67	2	*
	full NP	-.17	.08	-2.16	1.18	.06	18.13			
surprisal		.05	.04	1.46	.23	.03	7.80	75.18	2	*

Table C.7: Two Multinomial logit models predicting mention type (baseline level is “pronoun”), based on 1) surprisal alone and 2) shallow linguistic features + surprisal. Chi-square values of likelihood-ratio tests are indicative of any significant improvement in model by adding a predictor. * : $p_{\chi^2} < 0.001$.

		β	s.e.	t	p_t	F	df	p_F
Intercept		1.87	.02	80.75	$*$			
surprisal		0.25	0.02	10.74	$*$	115.32	1	$*$
Intercept		1.81	.05	40.08	$*$			
distance		0.17	0.02	7.08	$*$	50.08	1	$*$
frequency		-.13	.02	-5.37	$*$	28.82	1	$*$
antecedent	previous subject	-.51	.06	-8.46	$*$	71.59	1	$*$
mention	subject	.04	.05	.77	.44	.60	1	.44
antecedent type	proper name	-.21	.06	-3.21	.0013	59.32	2	$*$
	full NP	.42	.06	7.51	$*$			
surprisal		.17	.02	7.37	$*$	54.37	1	$*$

Table C.8: Two Linear regression models predicting mention length (number of tokens) of the masked mention, based on 1) surprisal alone and 2) shallow linguistic features + surprisal. F-test compares the fits of nested models. All predictors were tested to improve goodness-of-fit to the data except ”target mention is subject”. $*: p < 0.001$

Appendix D

MORE INFORMATION ABOUT THE META-ANALYSIS

D.1. Estimated effect size from individual comparison pairs

In the first stage of our meta-analysis, we computed summary statistics for each comparison pair (more predictable referents vs. less predictable referents). We did this separately for subject referents (e.g., “Paul” in “Paul embarrassed Alan because Paul...” vs. “Paul” in “Paul liked Alan because Paul...”) and object referents (e.g., “Alan” in “Paul liked Alan because Alan...” vs. “Alan” in “Paul embarrassed Alan because Alan...”) to control for the well-established effects of grammatical roles on pronoun production.

The effect of each individual comparison pair was estimated using Bayesian logistic regression models, in which pronoun use was predicted by referent predictability (binary factor; whether a referent is more or less predictable in a pair). For pro-drop languages, we constructed separate models for overt pronouns (use of overt pronouns predicted by whether a referent is more or less predictable in a pair) and null pronouns (use of null pronouns predicted by whether a referent is more or less predictable in a comparison pair). We extracted posterior draws from the regression models and presented the means and 95% credible intervals in Table D.1 for subject referents, and Table D.2 for object referents.

Experiment	Year	Language	Pronoun Type	Estimate	Error	95% CIs
Arnold	2001	English	pronoun/null	0.12	0.64	[-1.17, 1.36]
Contemori and Di Domenico (Exp.2)	2021	Italian	pronoun/null	-0.37	1.27	[-3.19, 1.83]
Contemori and Di Domenico (Exp.2)	2021	Italian	overt	-0.84	1.55	[-3.88, 2.29]
Contemori and Di Domenico (Exp.1)	2021	Spanish	pronoun/null	-0.60	0.66	[-2.06, 0.56]
Contemori and Di Domenico (Exp.1)	2021	Spanish	overt	0.62	0.66	[-0.52, 2.06]

Fukumura and van Gompel (Exp.1)	2010	English	pronoun/null	0.04	0.25	[-0.44, 0.53]
Fukumura and van Gompel (Pre.1)	2010	English	pronoun/null	-0.72	0.67	[-2.20, 0.44]
Holler and Suckow (ICV2 + because/since vs. but/although)	2016	German	pronoun/null	0.37	0.23	[-0.08, 0.83]
Holler and Suckow (ICV + because/since)	2016	German	pronoun/null	0.44	0.24	[-0.03, 0.92]
Holler and Suckow (ICV1 + because/since vs. but/although)	2016	German	pronoun/null	0.08	0.26	[-0.44, 0.59]
Hwang et al. (ICV + because)	2022	Mandarin	pronoun/null	0.12	1.39	[-2.06, 3.27]
Hwang et al. (ICV + because)	2022	Mandarin	overt	-0.43	0.38	[-1.16, 0.32]
Hwang et al. (ICV1 + because vs. so)	2022	Mandarin	pronoun/null	-0.93	0.76	[-2.34, 0.61]
Hwang et al. (ICV1 + because vs. so)	2022	Mandarin	overt	-0.33	0.32	[-0.93, 0.29]
Hwang et al. (ICV2 + because vs. so)	2022	Mandarin	pronoun/null	1.11	1.30	[-0.83, 4.17]
Hwang et al/ (ICV2 + because vs. so)	2022	Mandarin	overt	0.13	0.38	[-0.59, 0.87]
Hwang et al. (TPV + so)	2022	Mandarin	pronoun/null	1.82	1.17	[-0.01, 4.39]
Hwang et al. (TPV + so)	2022	Mandarin	overt	0.03	0.27	[-0.48, 0.57]
Hwang et al. (TPV1 + because vs. so)	2022	Mandarin	pronoun/null	0.66	0.50	[-0.26, 1.71]
Hwang et al. (TPV1 + because vs. so)	2022	Mandarin	overt	0.05	0.19	[-0.32, 0.41]
Hwang et al. (TPV2 + because vs. so)	2022	Mandarin	pronoun/null	2.04	1.18	[0.16, 4.77]
Hwang et al. (TPV2 + because vs. so)	2022	Mandarin	overt	0.24	0.29	[-0.32, 0.82]
Hwang continuity	2022	Korean	pronoun/null	0.96	0.13	[0.70, 1.22]
Hwang marker (Exp1 ICV)	2022	Korean	pronoun/null	0.59	0.39	[-0.16, 1.39]
Hwang marker (Exp1 TPV)	2022	Korean	pronoun/null	-0.08	0.48	[-1.03, 0.83]
Hwang marker (Exp2 ICV)	2022	Korean	pronoun/null	-1.48	0.40	[-2.24, -0.66]
Hwang marker (Exp2 TPV)	2022	Korean	pronoun/null	-0.31	0.38	[-1.03, 0.49]
Konuk and von Heusinger	2021	Turkish	pronoun/null	1.87	0.22	[1.43, 2.32]
Konuk and von Heusinger	2021	Turkish	overt	-1.31	0.48	[-2.29, -0.40]
Lam and Hwang	2022	Mandarin	pronoun/null	-0.29	0.14	[-0.56, -0.03]
Lam and Hwang	2022	Mandarin	overt	0.01	0.18	[-0.34, 0.38]
Liao	2022	English	pronoun/null	-0.04	0.13	[-0.29, 0.22]
Mayol	2018	Catalan	pronoun/null	-0.38	0.42	[-1.24, 0.39]
Mayol	2018	Catalan	overt	-0.03	0.63	[-1.20, 1.25]
Medina Fetterman et al. (spoken, different gender)	2022	Spanish	pronoun/null	0.04	0.38	[-0.71, 0.76]
Medina Fetterman et al. (spoken, same gender)	2022	Spanish	pronoun/null	0.92	0.38	[0.18, 1.68]
Medina Fetterman et al. (spoken, different gender)	2022	Spanish	overt	0.35	0.51	[-0.63, 1.38]
Medina Fetterman et al. (spoken, same gender)	2022	Spanish	overt	1.04	0.65	[-0.17, 2.39]
Medina Fetterman et al. (written, different gender)	2022	Spanish	pronoun/null	-0.26	0.27	[-0.79, 0.27]
Medina Fetterman et al. (written, same gender)	2022	Spanish	pronoun/null	0.76	0.27	[0.20, 1.30]
Medina Fetterman et al. (written, different gender)	2022	Spanish	overt	-1.13	0.54	[-2.22, -0.14]
Medina Fetterman et al. (written, same gender)	2022	Spanish	overt	-0.55	0.68	[-1.93, 0.73]
Patterson et al.	2022	German	pronoun/null	-9.46	9.80	[-36.59, -0.89]
Rohde and Kelher	2014	English	pronoun/null	-0.18	0.38	[-0.95, 0.56]
Solstad and Bott (Exp.1, because)	2022	German	pronoun/null	1.85	1.46	[-1.43, 4.39]
Solstad and Bott (Exp.1, so)	2022	German	pronoun/null	3.39	1.33	[0.81, 6.08]
Solstad and Bott (Exp.2)	2022	German	pronoun/null	-0.35	0.28	[-0.92, 0.21]
Weatherford and Arnold (Exp.1)	2021	English	pronoun/null	0.18	0.19	[-0.19, 0.55]
Weatherford and Arnold (Exp.2)	2021	English	pronoun/null	0.32	0.22	[-0.09, 0.75]
Zerkle and Arnold	2019	English	pronoun/null	0.82	0.22	[0.39, 1.24]
Zhan et al.	2020	Mandarin	pronoun/null	-0.89	0.43	[-1.71, 0.00]
Zhan et al.	2020	Mandarin	overt	-0.26	0.33	[-0.95, 0.37]
Kehler and Rohde	2019	English	pronoun/null	-0.83	0.55	[-1.98, 0.19]

Portele and Bader	2020	German	pronoun/null	-0.71	0.50	[-1.75, 0.23]

Table D.1: Summary of the posterior distribution for each comparison pair of subject referents. The estimate represents the effect size (in log odds) for each comparison pair, estimated based on the data gathered from each experiment.

Experiment	Year	Language	Pronoun Type	Estimate	Error	95% CIs
Arnold	2001	English	pronoun/null	0.81	0.40	[0.04, 1.65]
Contemori and Di Domenico (Exp.2)	2021	Italian	pronoun/null	1.14	0.37	[0.41, 1.85]
Contemori and Di Domenico (Exp.2)	2021	Italian	overt	-1.18	0.37	[-1.92, -0.44]
Contemori and Di Domenico (Exp.1)	2021	Spanish	pronoun/null	-0.51	0.38	[-1.28, 0.20]
Contemori and Di Domenico (Exp.1)	2021	Spanish	overt	0.50	0.38	[-0.20, 1.25]
Fukumura and van Gompel (Exp.1)	2010	English	pronoun/null	-0.05	0.22	[-0.48, 0.38]
Fukumura and van Gompel (Pre.1)	2010	English	pronoun/null	0.98	0.29	[0.39, 1.55]
Holler and Suckow (ICV2 + because/since vs. but/although)	2016	German	pronoun/null	0.12	0.19	[-0.25, 0.49]
Holler and Suckow (ICV + because/since)	2016	German	pronoun/null	0.28	0.21	[-0.13, 0.68]
Holler and Suckow (ICV1 + because/since vs. but/although)	2016	German	pronoun/null	0.13	0.22	[-0.30, 0.56]
Hwang et al. (ICV + because)	2022	Mandarin	pronoun/null	30.40	35.92	[-0.42, 134.18]
Hwang et al. (ICV + because)	2022	Mandarin	overt	0.84	0.85	[-0.57, 2.77]
Hwang et al. (ICV1 + because vs. so)	2022	Mandarin	pronoun/null	39.61	61.32	[-11.71, 229.12]
Hwang et al. (ICV1 + because vs. so)	2022	Mandarin	overt	0.79	0.84	[-0.60, 2.66]
Hwang et al. (ICV2 + because vs. so)	2022	Mandarin	pronoun/null	-0.02	1.37	[-2.23, 3.09]
Hwang et al. (ICV2 + because vs. so)	2022	Mandarin	overt	2.42	1.21	[0.57, 5.21]
Hwang et al. (TPV + so)	2022	Mandarin	pronoun/null	-1.40	1.61	[-4.39, 1.99]
Hwang et al. (TPV + so)	2022	Mandarin	overt	1.65	1.26	[-0.27, 4.60]
Hwang et al. (TPV1 + because vs. so)	2022	Mandarin	pronoun/null	-0.07	1.41	[-2.51, 3.04]
Hwang et al. (TPV1 + because vs. so)	2022	Mandarin	overt	0.97	1.28	[-1.15, 3.80]
Hwang et al. (TPV2 + because vs. so)	2022	Mandarin	pronoun/null	-1.02	1.30	[-3.79, 1.39]
Hwang et al. (TPV2 + because vs. so)	2022	Mandarin	overt	0.10	0.37	[-0.64, 0.85]
Hwang marker (Exp.1 ICV)	2022	Korean	pronoun/null	7.42	9.14	[-1.31, 32.74]
Hwang marker (Exp.1 TPV)	2022	Korean	pronoun/null	-0.03	1.59	[-3.24, 3.13]
Hwang marker (Exp.2 ICV)	2022	Korean	pronoun/null	-0.59	0.71	[-1.88, 0.91]
Hwang marker (Exp.2 TPV)	2022	Korean	pronoun/null	49.96	96.79	[-1.26, 258.92]
Konuk and von Heusinger	2021	Turkish	pronoun/null	0.16	0.30	[-0.40, 0.75]
Konuk and von Heusinger	2021	Turkish	overt	-0.71	0.73	[-2.14, 0.75]
Mayol	2018	Catalan	pronoun/null	0.48	0.25	[-0.01, 0.97]
Mayol	2018	Catalan	overt	0.33	0.40	[-0.41, 1.16]
Medina Fetterman et al. (spoken, different gender)	2022	Spanish	pronoun/null	0.05	1.54	[-3.02, 3.06]
Medina Fetterman et al. (spoken, same gender)	2022	Spanish	pronoun/null	-1.59	1.37	[-4.73, 0.68]
Medina Fetterman et al. (spoken, different gender)	2022	Spanish	overt	0.14	0.41	[-0.64, 0.96]
Medina Fetterman et al. (spoken, same gender)	2022	Spanish	overt	1.84	0.61	[0.73, 3.14]
Medina Fetterman et al. (written, different gender)	2022	Spanish	pronoun/null	-0.67	0.53	[-1.73, 0.33]
Medina Fetterman et al. (written, same gender)	2022	Spanish	pronoun/null	-0.52	0.56	[-1.66, 0.56]
Medina Fetterman et al. (written, different gender)	2022	Spanish	overt	0.85	0.35	[0.15, 1.55]
Medina Fetterman et al. (written, same gender)	2022	Spanish	overt	0.88	0.46	[0.01, 1.82]
Patterson et al.	2022	German	pronoun/null	0.65	0.35	[-0.04, 1.33]
Rohde and Kelher	2014	English	pronoun/null	-0.24	0.37	[-0.94, 0.48]

Solstad and Bott (Exp.1, because)	2022	German	pronoun/null	0.78	0.61	[-0.48, 1.91]
Solstad and Bott (Exp.1, so)	2022	German	pronoun/null	0.27	0.36	[-0.46, 0.95]
Solstad and Bott (Exp.2)	2022	German	pronoun/null	0.09	0.21	[-0.35, 0.51]
Weatherford and Arnold (Exp.1)	2021	English	pronoun/null	0.95	0.19	[0.59, 1.33]
Weatherford and Arnold (Exp.2)	2021	English	pronoun/null	0.91	0.21	[0.50, 1.31]
Zerkle and Arnold	2019	English	pronoun/null	-0.23	0.36	[-0.93, 0.45]
Zhan et al.	2020	Mandarin	pronoun/null	0.90	1.26	[-1.11, 3.91]
Zhan et al.	2020	Mandarin	overt	0.63	0.33	[0.02, 1.28]
Kehler and Rohde	2019	English	pronoun/null	0.70	0.30	[0.10, 1.29]
Portele and Bader	2020	German	pronoun/null	1.08	0.47	[0.22, 2.06]

Table D.2: Summary of the posterior distribution for each comparison pair of object referents. The estimate represents the effect size (in log odds) for each comparison pair, estimated based on the data gathered from each experiment.

D.2. The impact of factors on the relationship between predictability and use of the most reduced form

The summary of the model can be found in Table 6.4 of Section 6.3.2. In terms of language families, the analysis shows very weak evidence for a comparatively larger effect of predictability on Turkish null pronouns with a $\beta = 0.92$, 95% credible interval of $[0.04, 1.87]$, and a high probability of β being greater than 0 (≈ 0.98). This finding emerges from our analysis; however, to our knowledge, the existing literature does not provide any hypotheses that may explain this observation. We also note that this result should be interpreted cautiously since the posterior distribution is very wide given that we only have one sample of Turkish (Konuk and von Heusinger, 2021) included in our analysis, as shown in Figure D.1(f). We encourage future research to further explore this subject by examining Turkish and the underlying factors contributing to this observation.

D.3. The impact of factors on predictability effects in pro-drop languages

The summary of the model can be found in Table 6.5 of Section 6.3.3. The analysis shows weak evidence that predictability has a smaller effect in implicit causality scenarios, consistent with the speculations put forth by Rosa and Arnold (2017): $\beta = -0.37$, 95% credible interval $= [-0.67, -0.07]$, with a high probability (≈ 0.99) of β being less than 0. In addition, there is also a marginal effect of an interaction between grammatical role and pronoun type; $\beta = 0.15$, 95%

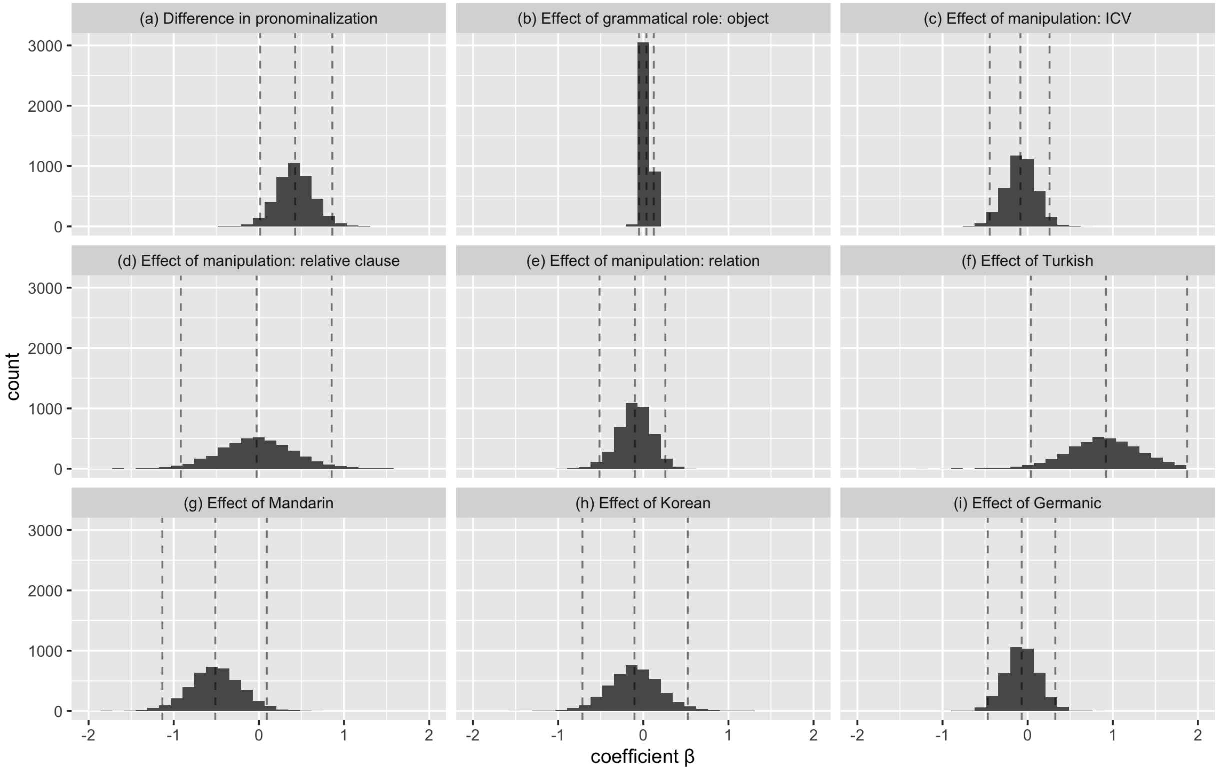


Figure D.1: Histograms displaying the posterior distribution for (a) the difference in the use of the most reduced referential form in a meta-analysis with covariates, and (b-i) the impact of covariates.

credible interval $= [0.03, 0.26]$, with a high probability of β being greater than 0 of ≈ 0.99. The effect could be due to there being were fewer occurrences of null pronouns than overt pronouns for object referents, making it more difficult to detect the effect, as suggested in previous research (e.g. Lam and Hwang 2022; Hwang et al. 2022).

D.4. Sensitivity analyses

D.4.1. Excluding the corpus study

In the subsequent analyses, we exclude the corpus study of Liao (2022) and focus only on the data obtained in psycholinguistic experiments to rule out the possibility that including the corpus study biases the results or drives the uncertainty.

We refit a basic random-effects model without covariates for the remaining

25 independent experiments (comprising 72 odds ratios). By analyzing the effect of predictability in this manner, we obtained an overall estimated odds ratio of 1.35 [1.07, 1.69, 95% CIs], indicating that the most reduced referential form is 1.35 times more likely to be used for more predictable referents compared to less predictable referents. This effect size is very close to the one reported in Sections 6.3.1 (1.33 [1.05, 1.63, 95% CIs]), where we included the corpus study.

In conclusion, the sensitivity analysis demonstrates that the inclusion of the corpus study Liao (2022) does not introduce biases into our results.

D.4.2. Excluding Spanish studies

In this section, we address concerns regarding the comparability of Spanish data from Contemori and Di Domenico (2021) and Medina Fetterman et al. (2022) to data from other Romance languages in our analyses (see main text for an explanation of these concerns). We exclude the Spanish data and reanalyze the dataset to examine the potential impact on our results and the associated uncertainty. To estimate the effect of predictability on the use of the most reduced form, we begin by refitting a basic random-effects model without covariates for the remaining 23 independent experiments (comprising 63 comparison pairs). This approach yields a pooled estimated odds ratio of 1.38 [1.08, 1.73, 95% CIs], which is again very similar to the estimate reported in Sections 6.3.1 (1.33 [1.05, 1.63, 95% CIs],) when including the Spanish data.

Next, we refit the model with the Spanish data excluded and incorporate (a) grammatical role, (b) manipulation type, and (c) language family as covariates to account for potential sources of variation. After adjusting for these factors, the effect estimate increases to 1.73 [1.09, 2.77, 95% CIs]. This larger CI indicates greater uncertainty. The summary of this model, presented in Table D.3, displays similar estimates to those in Table 6.4 when the Spanish data was included in the analysis. Also, there is still no strong evidence supporting the influence of grammatical role, manipulation type, or language family on the results.

Finally, we refit the model specifically for pro-drop languages without the Spanish data. The overall effect is estimated to be 1.63 [0.77, 3.42, 95% CIs], similar in magnitude but with much greater uncertainty than the original model reported in the main text (Table 6.4). The summary of the model is provided in Table D.4. Although the estimates are similar to those in the original model, the estimate for the Romance languages exhibits larger uncertainty due to the reduced sample size.

Overall, the sensitivity analysis demonstrates that the inclusion of Spanish data from Contemori and Di Domenico (2021) and Medina Fetterman et al. (2022) does not particularly impact or distort our results.

	Estimate	Estimated error	95% CI
Intercept	0.55	0.24	[0.09, 1.02]
manipulationType ICV	-0.12	0.19	[-0.49, 0.24]
manipulationType relativeClause	-0.04	0.48	[-0.98, 0.90]
manipulationType relation	-0.12	0.21	[-0.57, 0.27]
languageFamily Turkish	0.83	0.47	[-0.13, 1.74]
languageFamily Mandarin	-0.58	0.33	[-1.26, 0.06]
languageFamily Korean	-0.18	0.35	[-0.87, 0.51]
languageFamily Germanic	-0.19	0.22	[-0.64, 0.26]
grammaticalRole object	0.06	0.05	[-0.03, 0.16]

Table D.3: Effect estimate (in log odds) of predictability on the use of the most reduced referential form: Summary of the model with three covariates (manipulation type, language family, and grammatical role), with the Spanish data excluded.

	Estimate	Estimated error	95% CI
Intercept	0.49	0.37	[-0.26, 1.23]
manipulationType ICV	-0.40	0.18	[-0.76, -0.04]
manipulationType relation	0.18	0.15	[-0.12, 0.47]
languageFamily Turkish	0.71	0.81	[-0.87, 2.33]
languageFamily Mandarin	-0.24	0.55	[-1.38, 0.85]
languageFamily Korean	-0.41	0.59	[-1.58, 0.81]
grammaticalRole object	0.08	0.08	[-0.09, 0.24]
pronounType overt	-0.13	0.08	[-0.28, 0.02]
grammaticalRole object:pronounType overt	0.03	0.07	[-0.10, 0.16]

Table D.4: Effect estimate (in log odds) of predictability in pro-drop languages: summary of the model with four covariates added (manipulation type, language family, grammatical role, and pronoun type), with Spanish data excluded.

D.4.3. Excluding two studies that aimed at addressing other related questions

Contemori and Di Domenico (2021) and Solstad and Bott (2022) did not conduct experiments specifically aimed at addressing our research question. However, we included them in our analysis as they employed the same experimental paradigm as other studies and their data allowed for the calculation of the effect size of predictability. Here we exclude their data and reanalyze the dataset to examine the potential impact on our results and the associated uncertainty.

To estimate the effect of predictability on the use of the most reduced form, we begin by refitting a basic random-effects model without covariates for the remaining 22 independent experiments (comprising 63 comparison pairs). This approach yields a pooled estimated odds ratio of 1.34 [1.04, 1.68, 95% CIs], which is very close to the estimate reported in Sections 6.3.1 (1.33 [1.05, 1.63, 95% CIs]) when including the data from these two studies.

Next, we refit the model with these two studies excluded and incorporate (a) grammatical role, (b) manipulation type, and (c) language family as covariates to account for potential sources of variation. After adjusting for these factors, the effect estimate increases to 1.52 [0.98, 2.39, 95% CIs], indicating greater uncertainty. The summary of this model, presented in Table D.5, displays similar estimates to those in Table 6.4 when these two studies were included in the analysis. Also, there is still no strong evidence supporting the influence of grammatical role, manipulation type, or language family on the results.

Finally, we refit the model specifically for pro-drop languages without these two studies. The overall effect is estimated to be 1.58 [0.82, 3.03, 95% CIs], with great uncertainty. The model summary is provided in Table D.6. The estimates are very similar to those in Table 6.4.

Overall, the sensitivity analyses demonstrate that the inclusion of data from Contemori and Di Domenico (2021) and Solstad and Bott (2022) does not particularly impact or distort our results.

D.4.4. Using individual languages as a covariate instead of language families

In this section, we present a sensitivity analysis examining the impact of using individual languages as opposed to language families as a covariate. When incorporating language family, manipulation type and grammatical role as covariates, the summary estimate of the effect of predictability on the most reduced reference form is 1.54 [1.01, 2.36]. When using individual languages instead of language families as a covariate, the estimate increases to 1.65 [1.06, 2.61], accompanied by a slightly higher degree of uncertainty (see Table D.7).

	Estimate	Estimated Error	95% CI
Intercept	0.42	0.22	[-0.02, 0.87]
manipulationType ICV	-0.09	0.19	[-0.48, 0.26]
manipulationType relative Clause	-0.03	0.46	[-0.93, 0.88]
manipulationType relation	-0.10	0.20	[-0.51, 0.30]
languageFamily Turkish	0.93	0.48	[-0.03, 1.89]
languageFamily Mandarin	-0.52	0.32	[-1.15, 0.09]
languageFamily Korean	-0.09	0.34	[-0.77, 0.63]
languageFamily Germanic	-0.05	0.21	[-0.49, 0.37]
grammaticalRole object	0.03	0.05	[-0.07, 0.12]

Table D.5: Effect estimate (in log odds) of predictability on the use of the most reduced reference form: summary of the model with three covariates added (manipulation type, language family, and grammatical role), with Contemori and Di Domenico (2021) and Solstad and Bott (2022) excluded.

	Estimate	Estimated error	95% CI
Intercept	0.46	0.33	[-0.20, 1.11]
manipulationType ICV	-0.35	0.18	[-0.72, 0.00]
manipulationType relation	0.18	0.15	[-0.12, 0.46]
languageFamily Turkish	0.80	0.69	[-0.61, 2.20]
languageFamily Mandarin	-0.31	0.47	[-1.26, 0.72]
languageFamily Korean	-0.35	0.53	[-1.43, 0.77]
grammaticalRole object	0.06	0.07	[-0.08, 0.19]
pronounType overt	0.10	0.07	[-0.04, 0.25]
grammaticalRole object:pronounType overt	0.24	0.07	[0.12, 0.37]

Table D.6: Effect estimate (in log odds) of predictability in pro-drop languages: summary of the model with four covariates added (manipulation type, language family, grammatical role, and pronoun type), with Contemori and Di Domenico (2021) and Solstad and Bott (2022) excluded.

Furthermore, the effect of predictability in pro-drop languages is estimated to be 1.54 [0.80, 2.92], when individual languages are utilized as a covariate instead of language families (refer to Table D.8), which also exhibits much increased uncertainty. Estimates on individual languages for which only a single study is available are particularly uncertain (e.g. for Catalan, Italian).

In conclusion, utilizing individual languages rather than language families as a covariate in the analysis leads to increased uncertainty in the results, but does not particularly alter our results.

D.5. Exploratory analysis: the influence of task modality on the relationship between referent predictability and pronoun use

In a recent study by Ye and Arnold (2023), the impact of predictability on pronoun usage was found to be dependent on task modality. The study reported an influence of implicit causality in spoken tasks, but not in written tasks, suggesting that the effect of predictability is more pronounced in interactive communicative contexts.

We conducted an exploratory analysis with task modality (written vs. spoken) as the sole covariate, restricting our examination to English and Spanish, the languages for which both spoken and written experimental data were available (8 studies, 11 experiments). Task modality was centered to ensure that the coefficients in the model represented deviations from the mean for each level.

By incorporating modality as a main-effect variable, the overall effect of predictability was estimated to be 1.35 [1.04, 1.73, 95% CIs]. The resulting model summary is presented in Table D.9. Our analysis offers weak evidence suggesting that the influence of predictability on the most reduced form might be less pronounced in written experiments compared to spoken ones, corroborating the findings by Ye and Arnold (2023). The effect in spoken tasks is estimated to be above 0 (=1 in odds ratio) while the estimated effect in written tasks is around 0. However, the available evidence in our analysis is insufficient to draw definitive conclusions.

	Estimate	Estimated error	95% CI
Intercept	0.50	0.23	[0.06, 0.96]
manipulationType ICV	-0.15	0.19	[-0.53, 0.23]
manipulationType relativeClause	-0.02	0.48	[-0.99, 0.97]
manipulationType relation	-0.15	0.22	[-0.59, 0.26]
language German	-0.06	0.30	[-0.62, 0.55]
language Mandarin	-0.53	0.35	[-1.22, 0.17]
language Catalan	-0.10	0.52	[-1.14, 0.95]
language Spanish	-0.67	0.37	[-1.43, 0.08]
language Turkish	0.92	0.52	[-0.10, 1.96]
language Korean	-0.14	0.36	[-0.85, 0.59]
language English	-0.14	0.26	[-0.66, 0.36]
grammaticalRole object	0.03	0.05	[-0.06, 0.12]

Table D.7: Effect estimate (in log odds) of predictability on the use of the most reduced reference form: Summary of the model with manipulation type, language, and grammatical role added as covariates.

	Estimate	Estimated Error	95% CI
Intercept	0.43	0.31	[-0.22, 1.07]
manipulationType ICV	-0.39	0.17	[-0.72, -0.05]
manipulationType relation	0.19	0.15	[-0.11, 0.47]
language Mandarin	-0.21	0.49	[-1.18, 0.81]
language Catalan	0.21	0.82	[-1.48, 1.85]
language Spanish	-0.25	0.50	[-1.24, 0.76]
language Turkish	0.81	0.75	[-0.68, 2.28]
language Korean	-0.33	0.57	[-1.51, 0.83]
grammaticalRole object	0.07	0.07	[-0.05, 0.20]
pronounType overt	0.01	0.06	[-0.11, 0.14]
grammaticalRole object:pronounType overt	0.15	0.06	[0.03, 0.26]

Table D.8: Effect estimate (in log odds) of predictability in pro-drop languages: summary of the model with manipulation type, language, grammatical role, and pronoun type added as covariates.

	Estimate	Estimated Error	95% CI
Intercept	0.30	0.13	[0.04, 0.55]
modality written	-0.24	0.11	[-0.46, -0.01]

Table D.9: Effect estimate (in log odds) of predictability on the use of the most reduced reference form in English and Spanish: Summary of the model with task modality added as the only covariate.