ON THE LIMITATIONS OF DATA: MISMATCHES BETWEEN NEURAL MODELS OF LANGUAGE AND HUMANS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Forrest Lindberg Davis

August 2022

ON THE LIMITATIONS OF DATA: MISMATCHES BETWEEN NEURAL MODELS OF LANGUAGE AND HUMANS

Forrest Lindberg Davis, Ph.D.

Cornell University 2022

The majority of work at the intersection of computational linguistics and natural language processing aims to show, process by process, that human linguistic behavior (and knowledge) is reducible to a simple learning objective (e.g., predicting the next word) applied to unstructured linguistic data (e.g., written data). This dissertation uses three test cases to show concrete instances where current reductionist approaches fall short of human linguistic knowledge.

In the first case study, implicit causality, competition among multiple linguistic processes is shown to obscure human-like behavior in models. This challenges existing methodologies that rely on the investigation of individual linguistic processes in isolation and points to a mismatch between human linguistic systems and those built solely on the basis of linguistic data. In the second case study, ambiguous relative clause attachment, models of Spanish and English are compared to show that, while models appear to mimic humans in English, they fail to do so in Spanish. The failure of computational models of Spanish follows from a mismatch between data produced by speakers and speakers' interpretation preferences, and it is argued that this reflects fundamental limitations of text data. In the third case study, Principle B and incremental processing, it is demonstrated that, while humans use hard constraints to restrict their online processing of pronouns, computational models do not. The inability of models to process language incrementally like humans indicates a mismatch between linguistic data and the human parser.

This dissertation argues that data are not sufficient to instruct models about fundamental aspects of human language. Ultimately, in using techniques from psycholinguistics and careful cross-linguistic comparison, it is argued that neural models can reveal specific areas of linguistic knowledge where data are not enough, suggesting in turn what the human mind itself must contribute.

BIOGRAPHICAL SKETCH

Forrest Davis was born in Tampa, Florida and raised in Hailey, Idaho. He received his B.A. from Columbia University with a major in Computer Science and Mathematics. He is graduating from Cornell University with a Ph.D. in Linguistics and a graduate minor in Cognitive Science. Starting this fall, he will be a Postdoctoral Associate in the Department of Linguistics and Philosophy at the Massachusetts Institute of Technology.

To the ones I love: Rachel, Dan, Lindi, and Holliann

Oh you're doing it wrong, dissecting the bird

Trying to find the song

(John Craigie, *Dissect the Bird*)

ACKNOWLEDGEMENTS

While the dissertation marks the end of my graduate education, it is also the culmination of nearly 30 years of school. Many people have provided critical support, mentorship, encouragement, friendship and love throughout these years, without which I certainly would not have made it to this point. Unfortunately, space precludes acknowledging all such individuals, so, in what follows, I focus on the people who have had an outsized impact on the last five years of my life.

Custom dictates that I begin with my dissertation committee: Marten van Schijndel, Dorit Abusch, Miloje Despić, and John Whitman. I extend my deepest gratitude to my advisor, Marten van Schijndel, without whom this dissertation would not have been possible. While I began working with him later in my graduate studies, Marty has had an huge influence on me as a researcher and as an individual. He has provided me with invaluable advice in research, presentations, writing, and, more broadly, in how to be a scientist.

I extend my warmest gratitude to Dorit Abusch, who has been an outstanding and deeply thoughtful mentor over my entire time at Cornell. From my first course with her my first semester of graduate school until now, Dorit has truly inspired my interest in linguistics and has consistently supported me in all areas of my life. Dorit has always pushed me and my work, for which I am deeply grateful. I would also like to thank Miloje Despić who sparked my interest in theoretical syntax and whose classes taught me how to think deeply and passionately about linguistics. Chats with Miloje have been one of the highlights of my time in graduate school. Finally, I would like to thank John Whitman for thoughtful comments and critiques on my work throughout graduate school. John has a tremendous ability to ask probing questions, and I have benefited immensely from conversations with him.

Starting with Phonology I my first semester, Abby Cohn has taught me how

to connect seemingly disparate literature. Throughout graduate school, I have benefited immensely from Abby's mentorship, guidance, and collaboration. Abby has always encouraged me to read widely and shared books with me many times, for which I am extremely grateful.

I would also like to extend my thanks to Gerry Altmann. Our serendipitous meeting in Russia, facilitated by the wonderful Yanina Prystauka, was one of the best moments of the last five years. Gerry has been an extremely generous, kind, and thoughtful mentor and friend to me. Chats with him have been the highlight of many of my weeks, and I look forward to many more.

Thanks to the NLP group at Cornell, including Yoav Artzi, Claire Cardie, Cristian Danescu-Niculescu-Mizil, Lillian Lee, David Mimno, Sasha Rush, Maria Antoniak, Jonathan Chang, Ana Smith, and Laure Thompson. The Cornell NLP group has been a kind and thoughtful space to present research, learn about recent papers, and workshop ongoing work.

I am grateful for the administrators and staff in the Cornell Linguistics Department. In particular, Gretchen Ryan and Jenny Tindall have always extended their help to me and to everyone in the department. I would like to thank them for their kindness over the years.

Thank you to my cohort-mates and colleagues in the Cornell Linguistics Department, including Andrea, Binna, Dan, Eszter, Francesco, Jacob, John, Joseph, Mia, Naomi, Rachel, Seung-Eun, Shohini, and Siree. I will always treasure my friendship with Joseph who has been an incredible friend. Frequent writing sessions, discussions, practice talks, and dinners with Mia and Rachel have been a source of joy throughout graduate school. I am deeply grateful to Mia for our chats and close readings of Chomsky's works. Additionally, I was sustained, over the last five years, in the morning by Gimme! Coffee and in the evening by the Rhine House,

which I frequented with Joseph and Rachel.

Thank you to Irene Vogel. Since our first meeting at the LSA in New York, Irene has steadfastly supported me in my academic work, career goals, and my personal life. I have and continue to greatly benefit from her kindness and wisdom.

It is impossible to put into words my love and gratitude for Rachel. Meeting her was the best part, by far, of my time in graduate school, and I cherish all the time we spend together. Rachel has provided me boundless kindness, encouragement, and support, throughout the years. Our daily, and sometimes hourly, chats are an endless source of delight and joy, regardless of the topic. My research, mental health, and life have been truly enriched by her in every way, and I will be forever grateful to her.

Finally, I would like to thank my parents, Dan and Lindi, and my sister, Holliann, without whom my life would be far worse, and my cat, Fig. They have always provided critical love and support, and I cannot express how much they all mean to me. Simply put, I would never have completed this journey, and many others, without them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The primary focus of this thesis is the relationship between linguistic knowledge and linguistic data. In the study of human language, this relationship falls under the question of "the poverty of the stimulus" (alternatively, "Plato's Problem"; see Chomsky, 1980, 1986). While this thesis contributes to that discussion, primarily in Chapter 6, its main focus is on what use neural models of language have for the scientific study of human linguistic knowledge.[1] In particular, I take mismatches between the linguistic-like systems learned by neural models and the linguistic systems exhibited by humans as evidence for a disconnect between the properties of linguistic data and human linguistic knowledge. Rather than asking about what biases (or a priori knowledge) an acquisition device **must** have in order to develop a human-like linguistic system, this approach focuses on the computational models from natural language processing themselves, treating them as models dominated by the biases (linguistic and otherwise) in data. Put another way, this work is not interested in human language acquisition or whether transformational generative grammar is the "correct" model of linguistic knowledge, but instead, is interested in what systems follow directly from linguistic data.

A focus on what aspects of linguistic knowledge neural models capture is a typical approach in evaluating their capacities. The "hype" around these models (and artificial intelligence more generally), follows, at least partially, from the fact that neural models are "naive" (i.e. not informed by certain theoretical commitments). When neural models of language pattern like humans, the results are considered

[1] In what follows, I used *neural models of language* to denote the class of models investigated, rather than *neural language model* which has a technical meaning that does not apply to all models (e.g., BERT)

interesting because it is assumed (errorfully) that linguistic theories can be made less rich (e.g., we can remove some aspect of "UG"). This thesis tests and ultimately challenges this dominant view via concrete examples. The origin of the attested mismatches between neural models and humans extends beyond particular model architectures (e.g., auto-regressive vs. bi-directional transformer language models), and, I argue, follows from general properties of linguistic data. Therefore, these results pose challenges to any computational model which proceeds by prioritizing data (to the exclusion of meaningful constraints on model representations, learning procedures, or data).

While neural models of language have shown overlap with human linguistic behavior (e.g., Warstadt et al., 2020a; Hu et al., 2020a), the vast majority of these claims follow from experiments which are conducted on only English (for discussion of the English bias within NLP more broadly see Bender, 2009; Mielke, 2016 and for notable exceptions see Ravfogel et al., 2018; Gulordava et al., 2018; Muller et al., 2021), or target linguistic processes in isolation (e.g., the checklist approach; see Ribeiro et al., 2020). In contrast, an important aspect of human linguistic knowledge is that comparable states of knowledge are obtained regardless of the specific language and involves an intricate system of linguistic processes and levels of representation. Theory development within transformational generative grammar, for example, proceeds via comparison of diverse linguistic systems and by situating a given process within the broader architecture of the human linguistic system (e.g., morphology, syntax, semantics). It is difficult, then, to compare the capacity of neural models and humans because neural models are tested in a much more narrow sense than is common in linguistic theory.

This thesis attempts to address the gap between claims in natural language pro-

cessing and the linguistic system in humans via careful cross-linguistic comparison and attention to the interaction of linguistic processes (both within neural models and humans). In particular, three phenomena are explored: implicit causality, ambiguous relative clause attachment, and binding principles and coreference processing. Ultimately, I find evidence that the linguistic knowledge of neural models is strongly dependent on the particular language under investigation, and is not robust to interactions with other linguistic processes. Therefore, whatever linguistic knowledge models have remains far from human linguistic knowledge.

Careful consideration of these limitations of neural models in capturing these three phenomena suggests three broader classes of mismatches between linguistic data and linguistic knowledge:

1. mismatches in constraint ranking
2. mismatches in production
3. mismatches in processing constraints

While resolution of these mismatches lies outside the scope of this thesis, these results suggests potentially fruitful areas where human linguistic knowledge (and, in turn, our linguistic theories) must extend beyond superficial properties of language data. In what follows I detail some existing literature, outline common techniques for evaluating neural models of language, sketch out the empirical bias for the at-issue human linguistic processes, and lay out a roadmap of the remaining chapters.

1.1 Background

Neural models of language follow from a rich history of using connectionist models as models of human cognition. Such approaches have enjoyed a steady popularity in psycholinguistic research since the 1980s (seminal work includes Rumelhart and McClelland, 1986; Elman, 1990).[2] The recent successes of large neural language models in NLP have inspired renewed interest in the relationship between human linguistic representations in the brain and neural model representations (e.g., Schrimpf et al., 2020; Goldstein et al., 2020; Heilbron et al., 2020).

Comparisons between neural models and human linguistic behavior (e.g., acceptabilities judgments, reading times) have been more widely undertaken. Stemming from Linzen et al. (2016), there has been a growing body of literature within computational linguistics focusing on the ability of neural language models to match human-like subject-verb agreement patterns (e.g., Enguehard et al., 2017; Bernardy and Lappin, 2017; Gulordava et al., 2018; Linzen and Leonard, 2018; Wilcox et al., 2018; McCoy et al., 2018; Giulianelli et al., 2018; Ravfogel et al., 2018, 2019; Wilcox et al., 2019b; An et al., 2019; Mueller et al., 2020; Wilcox et al., 2020b; Arehalli and Linzen, 2020; Warstadt et al., 2020a).

In addition to canonical subject-verb agreement, a number of other syntactic structures have been investigated. Neural language models have been claimed to exhibit human-like behavior in processing center embedding, syntactic islands (Wilcox et al., 2018, 2019b), and garden path constructions (van Schijndel and Linzen, 2018a; Futrell et al., 2018b; Frank and Hoeks, 2019). Additional syntactic structures that have been explored include anaphoric binding (Marvin and Linzen,

[2]See Altmann (2013) for a thorough survey of advances in the field of psycholinguistics as it relates to connectionism, context, and levels of linguistic representation.

2018; Futrell et al., 2018b; Warstadt et al., 2019b), negative polarity items (Marvin and Linzen, 2018; Futrell et al., 2018b; Jumelet and Hupkes, 2018; Warstadt et al., 2019a), and filler-gap dependencies (Chowdhury and Zamparelli, 2018, 2019; Da Costa and Chaves, 2020; Bhattacharya and van Schijndel, 2020). Moving beyond pure syntactic knowledge, referential knowledge acquired by neural language models has some degree of prominence in the recent literature (Clark et al., 2019; Sorodoc et al., 2020; Upadhye et al., 2020). Additionally, neural language models exhibit at least some facility with pragmatic and discourse structure (Jeretic et al., 2020; Schuster et al., 2020; Davis and van Schijndel, 2020b; Upadhye et al., 2020).

For the most part this work has focused on English, though there are some notable exceptions (e.g., Gulordava et al., 2018 looked at English, Italian, Hebrew, and Russian; Mueller et al., 2020 explored English, French, German, Russian, and Hebrew; An et al., 2019 compared French and English; and Ravfogel et al., 2018 focused on Basque). The present thesis addresses this gap by evaluating neural models for a variety of languages beyond English, including Spanish, Italian, and Chinese. Ultimately, I argue that cross-linguistic comparisons are critical for evaluating whether neural models can acquire human-like linguistic systems.

The results of the recent literature coupled with the lack of a strong prior for linguistic structure in the various models have led to claims that human-like linguistic structure can emerge solely from training on linguistic data. However, the linguistic representations of neural models are admittedly not as robust or general as humans and much of the existing literature suggests as much (e.g., van Schijndel et al., 2019; Bhattacharya and van Schijndel, 2020; Kodner and Gupta, 2020). Even in cases where there is overlap in human and model linguistic behaviors, there remains quantitative differences in effect sizes (see van Schijndel and Linzen, 2021).

Pertinent to this work is the growing acknowledgement, on (semi-)theoretical grounds, that natural language comprehension is not possible given text alone. Arguments that speaker intent is crucially missing from current language models have been advanced (Bender and Koller, 2020); as well as claims that embodiment, perception, and social interaction are necessary components to build a truly human-like model of language (Bisk et al., 2020). Much of this work, however, implicitly assumes that human-like language form, that is grammatical surface structures, may still be learnable from just text. For this perspective, the key missing ingredient is a mapping between form and meaning which remains unspecified in current training regimes that rely on text data alone. This thesis makes a stronger claim: linguistic form does not follow directly from linguistic data.

Finally, as aptly pointed out in Pannitto and Herbelot (2020), theoretical assumptions inherited from generative linguistics have colored the interpretation of a given neural model's linguistic abilities. That is, we expect models to acquire abstract linguistic processes that apply in a variety of specific cases, so it is assumed that model success points towards model abstractions and model failure towards inability to infer the correct abstract structure. In contrast to expecting a model to "idealize syntactic structure as a separate and more abstract ability from the knowledge of statistical regularities or lexical co-occurrences", we may need an approach focusing on individual constructions (Pannitto and Herbelot, 2020 p. 166; see Madabushi et al., 2020 for similar views). While this world view is not adopted in this thesis, the point is well taken. In Chapter 2, I lay out the key assumptions leveraged in this dissertation so as to facilitate meaningful comparisons to alternative perspectives.

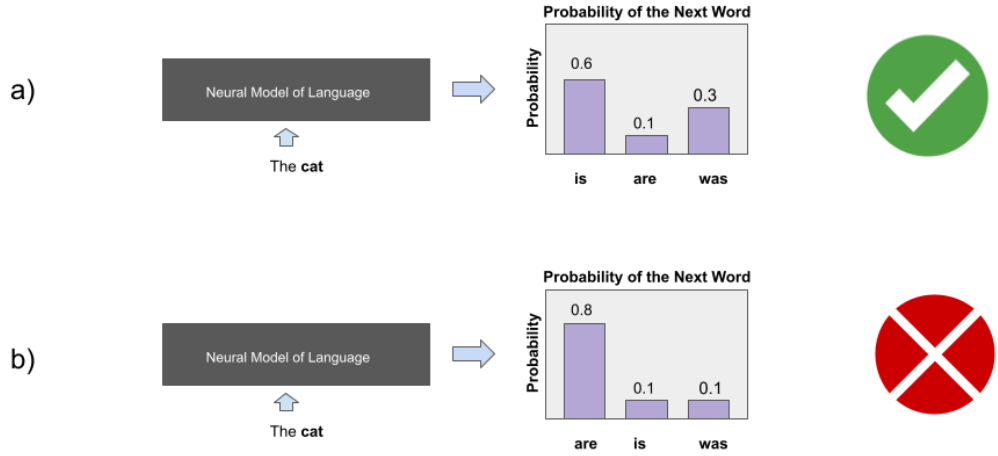

Figure 1.1: Schematic of targeted syntactic evaluations. The neural model assigns probability to the next words for the prefix *the cat*. In (a), we schematize a model which assigns more probability to singular verbs in its prediction, while in (b) we schematize a model which assigns more probability to plural verbs than singular verbs.

1.2 Techniques for Evaluating Neural Models of Language

A growing body of literature within natural language processing has coalesced around probing large, pre-trained neural models of language for aspects of human linguistic knowledge or structure. A number of methods of evaluating models have been proposed (for a review of some techniques, see Rogers et al., 2020), in this review I focus on three: targeted syntactic evaluations, adaptation (or fine-tuning), and representational probing.

1.2.1 Targeted Syntactic Evaluations

Targeted syntactic evaluations, as a methodology for evaluating neural models of language, proceeds along similar lines as minimal pairs in linguistics. The approach is schematized in Figure 1.1. Suppose we are interested in evaluating a neural model of language for knowledge of subject-verb agreement. We could construct minimal pairs like:

(1) a. The cat is hungry.
    b. *The cat are hungry.

In (1), the subject is singular but the agreeing verb differs in number. The general logic is that if a model has learned subject-verb agreement then its predictions about the upcoming verb should track grammaticality, where more probability should be assigned to singular verbs than plural verbs. In other words, (1-a) should be more probable than (1-b). Comparisons like this are aggregated across a number of minimal pairs. If model behavior consistently mimics humans (i.e. probability of grammatical sentences > probability of ungrammatical sentences), then we infer that this model has acquired subject-verb agreement.

Targeted syntactic evaluations for neural models were popularized by Linzen et al. (2016) for subject-verb agreement, but the approach has been expanded to other linguistic phenomena like islands, filler-gap dependencies, garden path sentences, negative polarity items, and reflexive pronouns (see Wilcox et al., 2018; Futrell et al., 2018b; van Schijndel and Linzen, 2018a; Wilcox et al., 2019a; Warstadt et al., 2019a; Bhattacharya and van Schijndel, 2020).

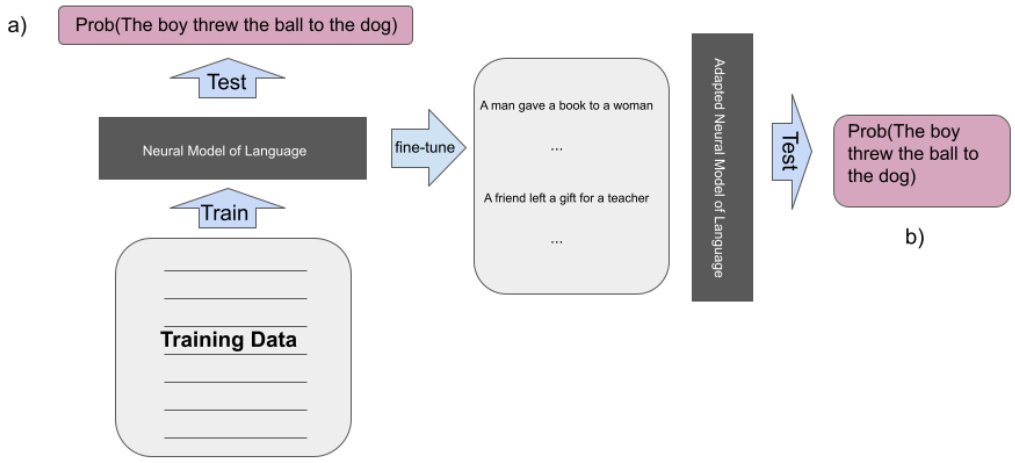

Figure 1.2: Schematic of the adaptation paradigm for evaluating the representation of the prepositional object construction for ditransitive verbs. The baseline neural model assigns some probability to examples of prepositional object constructions (a). This model is then adapted to non-overlapping examples of prepositional object constructions, and the probability of the same examples are calculated again (b). If the probability of (b) is greater than (a), we infer that the models have "primed" a (more general) representation of prepositional object constructions.

1.2.2 Adaptation

Adaptation in the literature on neural models of language can refer to priming a certain structure (e.g., expecting passives after seeing a passive) or to the adaptation of a model to a specific corpus (i.e. additional domain-specific training). Either use of adaptation relies on fine-tuning, training an already trained neural model on (a usually small amount of) additional data with certain properties (sketched in Figure 1.2).

Adaptation is often used as a means of understanding the underlying linguistic

knowledge of neural models (e.g., van Schijndel and Linzen, 2018b; Chowdhury and Zamparelli, 2019; Prasad et al., 2019). Changes in model behavior, conditioned on properties of the data used in adaptation, are taken as evidence that the model encodes the relevant property. For example, if adapting to instances of a certain type of relative clause results in a model which favors another type of relative clause, then it is claimed that these relative clause structures are abstractly related in the model.

1.2.3 Representational Probing

The above techniques for evaluating neural models make use of model behavior (i.e. the output of neural models). Other work evaluates the internal representation of neural models. This approach is now called *representational probing* (e.g., Ettinger et al., 2016; Belinkov et al., 2017; Hewitt and Liang, 2019) and makes use of what are called *probes* or *diagnostic classifiers* (e.g., Hupkes and Zuidema, 2018; Giulianelli et al., 2018). Comparing internal representations of connectionist models to human linguistic representations has a rich history beyond its current prevalence in natural language processing (e.g., Elman, 1991).

The typical approach, schematized in Figure 1.3, extracts the internal representations from a neural model of language and trains a classifier (e.g., a linear classifier; multi-layer perceptron) to predict a label corresponding to the linguistic feature of interest. For example, we may be interested in whether a neural model encodes part of speech information. To test this, we could train a classifier to predict from internal representations of words, their part of speech label. High accuracy of this classifier on held out data is taken as evidence that the model's internal representations encode part of speech information.

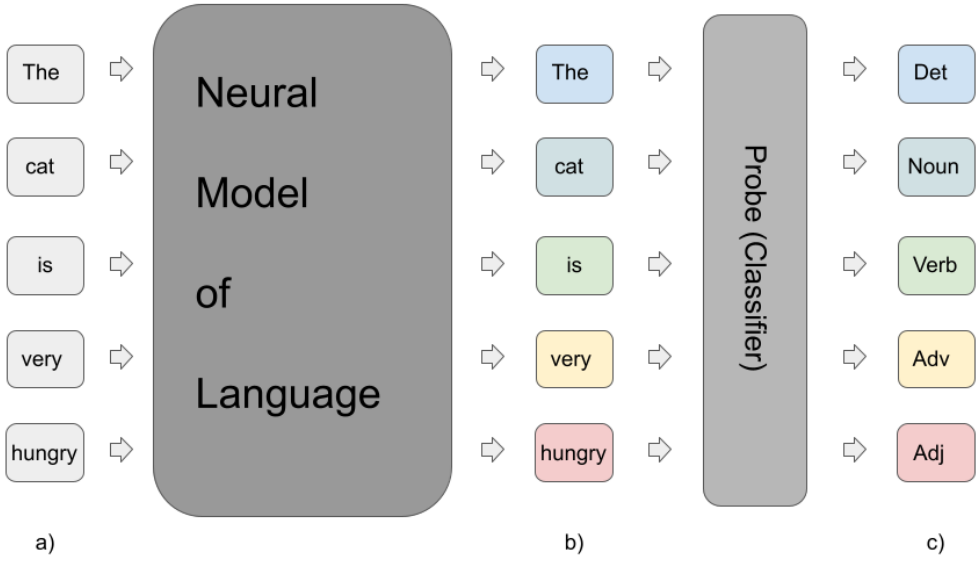

Figure 1.3: Schematic of the representational probing paradigm for uncovering linguistic representations in neural models of language, in this case part of speech information. Words are encoded by a neural model (a), and the resulting model internal representations of the words (b), are used to train a classifier (a probe) to predict part of speech labels (c). Accuracy of this probe (relative to some baseline) is used as a proxy for the presence of a corresponding representation of part of speech in the internal representation of the model.

1.3 Psycholinguistic Background for Thesis Experiments

This thesis focuses on three linguistic phenomena which are well studied in psycholinguistics: implicit causality, ambiguous relative clause attachment, and pronominal coreference (in relation to binding). Below I briefly sketch out each phenomenon and some literature relevant to this thesis.

1.3.1 Implicit Causality

The first phenomenon tested in this dissertation is implicit causality. Pronouns can exhibit ambiguity concerning their antecedent. Consider the following minimal pair:

(2) a. Sally frightened Keisha because she was so powerful.
b. Sally feared Keisha because she was so powerful.

In (2), *she* could refer to either *Sally* or *Keisha*.[3] However, English speakers preferentially interpret the pronoun as referring to *Sally* in (2-a) and to *Keisha* in (2-b). That is, certain verbs bias the interpretation of pronouns, with some having a "subject-bias" like *frightened* and others having a "object-bias" like *feared*. This phenomenon, called implicit causality, was noted in Garvey and Caramazza (1974), and has been an ongoing target of investigation in psycholinguistics (e.g., Kehler et al., 2007; Ferstl et al., 2011; Hartshorne and Snedeker, 2013; Hartshorne, 2014; Williams, 2020). Additionally, implicit causality has been attested in a number of languages other than English (for a review, see Hartshorne et al., 2013).

As it pertains to this thesis, implicit causality is claimed to be a linguistic process which does not rely on additional pragmatic inferences by comprehenders (e.g., Rohde et al., 2011; Hartshorne and Snedeker, 2013). Thus, implicit causality is argued to be contained within the linguistic signal, analogous to evidence for

[3]Of course, for particular persons named Sally or Keisha, *she* may not be their pronoun, and this would block reference to one or both of these individuals with the feminine pronoun *she*. I am not aware of a study in psycholinguistics that explores the relationship between implicit causality verb bias and gender identity. The results of such a study would extend work exploring the interaction of implicit causality and the relative status of the individuals mentioned in the event (Garvey et al., 1974), the animacy of the participants (Corrigan, 1988), and the valency of the event (Corrigan, 2001).

verb argument preferences and information about agreement. That is, we might reasonably expect neural models of language, which lack pragmatic competence, to learn an implicit causality bias. In Chapter 3, I investigated neural models of English, Chinese, Italian, and Spanish for IC biases, finding that competition between linguistic processes can obscure model knowledge.

1.3.2 Ambiguous Syntactic Attachment

The second phenomenon tested in this dissertation is ambiguous relative clause attachment. In a typical linguistics introductory course, students encounter the following stimulus as an example of syntactic ambiguity:

(3) The man ate the pizza with a fork.

The sentence in (3) is an example of prepositional phrase attachment ambiguity. Ignoring semantic plausibility, there are two possible syntactic parses for this sentence. One associating *with a fork* with the verb *ate* yielding an interpretation along the lines of "using a fork, the man ate the pizza". The other parse associates *with a fork* with the object *pizza* yielding an interpretation along the lines of "the man ate the pizza which had a fork on it". There are a number of syntactic ambiguities of this sort (for a useful compilation see Frazier and Clifton, 1996, p. 42-43).

As is clear from (3), the same surface string can correspond to different structures with different interpretations. In other words, there are not always surface cues of syntactic contrasts. In evaluating the behavior of neural models of language

using a technique like targeted syntactic evaluations (see Section 1.2.1), however, we need strings that are minimal different because models do not have a way of intending to mean something. This means that we cannot evaluate the prepositional phrase attachment ambiguity in models. In order to test ambiguous attachment in models, we must focus, instead, on cases in which surface contrasts can, in principle, distinguish the possible syntactic parses. One such case is ambiguous relative clause attachment.

(4) The man met the friend of the neighbor who is happy.

In (4), there are again two possible syntactic parses depending on the attachment location of the relative clause *who is happy* in the complex noun phrase *the friend of the neighbor*. In one, the relative clause attaches to the structurally higher noun *friend* yielding the interpretation that the friend is happy. In the other, the relative clause attaches to the structurally lower noun *neighbor* yielding the interpretation that the neighbor is happy.

It is well established that English speakers have a consistent preference in ambiguous relative clause attachment; they prefer attaching the relative clause to the lower noun (e.g., Carreiras and Clifton, 1993; Frazier and Clifton, 1996; Carreiras and Clifton, 1999; Fernández, 2003). Critically, this preference can be evidenced by constructing minimal pairs and measuring whether there is a difference in some measure of human processing. The high and low nouns can differ in number with agreement on the relative clause verb disambiguating the attachment location:

(5) a. The man met the **friend** of the neighbors who **is** happy.

b. The man met the friend of the **neighbors** who **are** happy.

In (5-a) the relative clause verb *is* agrees with the higher noun *friend*, while in (5-b) the relative clause verb *are* agrees with the lower noun *neighbors*. A number of studies using on-line measures have found that English speakers have slower reading times for stimuli like (5-a) as compared to (5-b) (e.g., Fernández, 2003).

A similar ambiguity can be constructed in other languages, where empirical evidence suggests that the English type lower noun preference is less common (Brysbaert and Mitchell, 1996). A proto-typical example of a language with a bias for attachment to the higher noun is Spanish (e.g., Carreiras and Clifton, 1993, 1999; Fernández, 2003). In Chapter 4, I compare English and Spanish neural models of language to investigate whether models trained on text can consistently capture human attachment preferences.

1.3.3 Pronominal Coreference and Binding Principles

The third, and final, phenomenon tested in this dissertation is the interaction between Principle B and incremental coreference processing. Factors beyond implicit causality, discussed above, influence the resolution of (potentially ambiguous) pronouns. Consider the following sentence from Chow et al. (2014):

(6) Bill explained to Mary that Peter had deceived him.

Despite *him* agreeing in gender with both *Peter* and *Bill*, *him* unambiguously refers to *Bill*. A property of grammar, Principle B from Binding Theory (Chomsky, 1981), blocks *him* from referring to *Peter*. Principle B, in simplified form states that a pronoun must be free within its local domain. If *him* was to refer to *Peter*, then

him would be bound in its local domain.

The competition between cues for pronominal antecedents (e.g., gender agreement) has been a major source of experimental work, and the cues themselves are often thought as constraints. Broadly, these constraints are binned into two classes: those that make reference to the agreement features on the pronoun (e.g., gender, number, person) and those that make reference to structural constraints (e.g., binding principles). A number of works have found that structural constraints immediately constrain the set of possible antecedents (e.g., Clifton et al., 1997; Sturt, 2003; Chow et al., 2014; Kush and Phillips, 2014; Kush and Dillon, 2021). Using (6) as an example, such work would suggest that the initial set of possible antecedents is {Bill, Mary}, with the gender of *him* excluding *Mary* later. However, other work has suggested that grammatically illicit antecedents can, in fact, have measurable effects (e.g., Badecker and Straub, 2002; Kennison, 2003). In other words the initial set may contain *Peter* as well. It may well be that there are task specific effects from the presentation modality (e.g., written vs auditory) that drive different candidate sets (as discussed in Nicol and Swinney, 2003), or that different measures capture different time points in processing, with later stages of processing potentially adding grammatically illicit candidates (Sturt, 2003). In Chapter 5, we explored whether neural models acquired syntactic constraints on coreference and whether their incremental processing behavior mirrored humans.

1.4 Roadmap

The rest of this dissertation is structured as follows. First, Chapter 2 lays out three basic assumptions necessary for interpreting neural models in this thesis. These are

(i) the relevant notion of language, (ii) the relationship between human linguistic processing and neural models, and (iii) the world-view (or theory) which mediates inferences from neural models to a broader understanding of human linguistic knowledge.

Chapter 3 investigates implicit causality. First, we find evidence that neural models of English capture aspects of both categorical and gradient implicit causality biases in pronoun production. However, when we look beyond English, we find that models fail to learn an IC bias in Spanish and Italian (but appear to do so for Chinese). We then link this failure to the competition between linguistic processes. Namely, pro-drop in Spanish and Italian obscures IC biases in neural models. Using adaptation, we show that demoting the competing constraint can surface otherwise dormant implicit causality behavior. Thus, while neural models may be able to acquire isolated linguistic processes, they struggle to arrive at human-like constraint rankings in cases where multiple linguistic processes target the same environment. The difference between neural models and humans suggests that linguistic data, itself, can lead models away from human-like generalizations.

Chapter 4 investigates ambiguous relative clause attachment. We show that while models may appear to mimic human attachment preferences in English, comparable models of Spanish fail to acquire a human-like attachment preference. Additionally, fine-grained investigation of English and Spanish demonstrates that neural models consistently over-emphasize a low attachment preference (when compared to humans) and fail to acquire interactions between attachment and other linguistic processes (i.e. implicit causality). The chapter concludes with an investigation of gender disambiguation for relative clause attachment in Spanish, providing evidence that neural models may be tracking initial parsing preferences

of Spanish speakers, rather than their ultimate interpretation preferences. Further inspection of the training data suggests that this mismatch is not driven entirely by model biases. Instead, the production biases that generate training data are not the same as the comprehension biases we ultimately want models to have. Thus, other data or modeling methods will be needed to resolve apparently inherent mismatches between production and comprehension.

Chapter 5 investigates the interaction between Principle B (Chomsky, 1981) and coreference behavior in neural models of English. We show that some neural models capture aspects of human behavior associated with Principle B. However, they fail to acquire a fuller range of Principle B interactions. Moreover, in comparing the incremental behavior of neural models with humans, we find that models do not pattern with humans in their consideration of pronominal antecedents. These results suggest that fundamental aspects of syntax and of human parsing mechanisms are not evidenced by data. Finally, Chapter 6 concludes by connecting the findings of this dissertation to arguments from poverty of the stimulus and discusses the broader connections between neural models and linguistic theory.

CHAPTER 2

ASSUMPTIONS IN THE STUDY OF NEURAL MODELS OF LANGUAGE

Behaviors (or more broadly empirical results) are interesting insofar as they are related to an explanatory theory. Put simply, results that are trivial are rarely interesting, and what we call trivial follows from our systems of explanation. This chapter proposes a set of assumptions necessary to determine what is interesting for work relating neural models of language to the scientific understanding of human linguistic capacities. I take the relevant assumptions to address the following three basic questions:

1. What is meant by the term language?
2. Which aspects of human linguistic processing are relevant?
3. What world view (or theory) mediates inferences from computational models to human capacities?

While clarifying the object of study, the relevant evaluation, and the underlying theory (i.e. the above basic questions) are crucial for understanding any empirical results, it is especially critical for relating neural models of language to Chomskyan linguistic theory because of the perceived difference in their evidentiary bases. Transformational generative linguistics has proceeded since its inception in the 1950s with the goal of carving out of the complexity of linguistic experience crucial data that "are *revelatory* of that [underlying] reality" (Rey, 2020, p. 17) which theories attempt to articulate.[1]

[1]I am indebted to the work in Rey (2020) which left its mark on this chapter and helped crystallize many of the ideas and issues I was having in interpreting the results from the experiments involving neural models both in subsequent chapters and in the field more broadly.

The relationship between data (i.e. empirical observations) and phenomena (i.e. the deep aspects of reality our theories attempt to uncover, following the terminology in Bogen and Woodward, 1988) is made more complex in generative linguistics because the capacity one seeks to characterize is the ability of humans to generate and interpret novel sentences (i.e. sentences they have not yet seen). Moreover, Chomskyan linguistic theory heightens the disconnect between observable data and (underlying) phenomena with the centrality of evidence from ungrammatical examples which definitionally should never exist. For example the sentence "Who did Anne and find Bill eating", under a modest degree of idealization away from the possibility of errorfully producing the sentence, will never be produced by a speaker.[2] This reliance on non-existing data has often been contrasted with more "empirically motivated" work in fields like natural language processing. Addressing the above questions is meant to carve out a meaningful basis for drawing on and bringing together insights from both fields.

In what follows, I consider both commonly assumed (but often unarticulated) responses to the above questions and lay out the specific positions that are assumed in the dissertation. Before addressing these three specific assumptions in Sections 2.2–2.4, I motivate them by discussing, in Section 2.1, two more general issues in reasoning from empirical results that currently pervade the field, which I title the *Error of Imprecise Expectations* and the *Error of Empirical Expansion*.

[2]As Rey (2020) notes these "WhyNots" are interesting within the study of language because they are often comprehensible upon reflection (in the case mentioned above, it is easy enough to get that the intended meaning is captured by the echo question "Anne and who found Bill eating"), yet are seemingly impossible to generate.

2.1 Some Errors in Interpreting Empirical Results

To facilitate understanding of the *Error of Imprecise Expectations* and the *Error of Empirical Expansion* it is useful to sketch out a typical methodological approach to studying computational models in natural language processing. While explicit explanatory theories are often not advanced in natural language processing, empirical results can be made informative via the creation of a *benchmark*, which serves as an evaluation metric, for a desired *domain of interest*, which picks out the relevant capacity being evaluated. Holding fixed the *benchmark* facilitates comparisons between disparate computational models by providing an evaluation procedure which marks progress towards modeling the *domain of interest*. Take for example, the subfield of natural language processing called natural language understanding (commonly referred to as NLU). Its *domain of interest* is language comprehension (the bounds of what is meant by comprehension are largely underspecified, which we return to in a moment), and the *benchmarks* pick out from the set of all possible comprehension behaviors some subset like answering reading comprehension questions after a passage (e.g., Lai et al., 2017) or labeling sentence pairs with relations like entailment (e.g., Bowman et al., 2015). Computational models are then compared according to their ability to perform on this benchmark (and are, moreover, compared to human performance on the benchmark). On the surface this approach seems tenable, though there are certainly problems in benchmark construction that go beyond the scope of the present discussion (for interesting discussion of some limitations see Bowman and Dahl, 2021).

There are two inferential errors commonly attested in the field which motivate this section. The first, the *Error of Imprecise Expectations*, deals with both imprecision in what criterion is used in evaluating a computational model and

imprecision in interpreting the degree of success of a computational model for a given criterion. Turning first to the imprecision in selecting the criterion, consider the computational model DALL-E (Ramesh et al., 2021). DALL-E is a model trained to generate images from a text prompt, and the images it produces are often considered impressive. The wide engagement with DALL-E makes it a useful test case for the *Error of Imprecise Expectations*, because there is a disconnect between the criteria that individuals are using to assess the results of DALL-E and the specific evaluation metrics utilized in the technical paper.

A number of news media articles presenting DALL-E claim that the images it creates are "startlingly accurate" (Business Insider), show "some of the same creativity that human cartoonists do" (New York Times), and are "entirely new" (in the words of Illya Sutskever, OpenAI's co-founder, in the Wall Street Journal).[3] The relevant criteria, then, that inform these interpretations of results from DALL-E include creativity, novelty, and accuracy. The quantitative human evaluations reported in the technical paper, however, compare outputs from DALL-E to outputs from a similar computational model along two dimensions: which is a better fit to the caption and which is more realistic (see Figure 13 in Ramesh et al., 2021).

That is, of the three criteria implicit in the news reports (and which seemingly guide evaluations of the success of DALL-E for the broader community of laypeople), only one, accuracy, was (to some extent) systematically investigated. Novelty, for example, is not explicitly mentioned in the text, however, their qualitative results suggest that the "model has the ability to generalize in ways that [they] did not originally anticipate" (Ramesh et al., 2021, p. 8). However, for the example caption given ("a tapir made of accordion"), there is no evaluation of the novelty of the

[3]See `https://www.businessinsider.com/dall-e-mini`, `https://www.nytimes.com/2023/04/15/magazine/ai-language.html`, and `https://www.wsj.com/articles/how-computers-with-humanlike-senses-will-change-our-lives-11625760066`.

caption, and therefore the novelty of the generated image (e.g., how often does the word 'tapir' or the phrase "made of" appear in the training captions).

The *Error of Imprecise Expectations*, then, is that the criteria that individuals use to assess the results of a neural model are varied and often difficult to formalize (e.g., how does one operationalize "creativity"?), and to the extent that a specific criterion can be extracted, it is not necessarily explicitly used to develop or formally evaluate the model. In the case of DALL-E, this imprecision is compounded by the fact that the training data are not available to the public or to the broader research community. If we take novelty to be one main component for assessing results from DALL-E, we cannot even begin evaluating the model with this criterion without access to the training data. This is a broader issue in cases where a given neural model exceeds peoples (implicit) expectations. Very little work explores the training data for such models, so it is difficult to systematically assess what constitutes reasonable behavior for the computational model. Thus the *Error of Imprecise Expectations* will remain.

Turning to the other case of the *Error of Imprecise Expectations*, consider again natural language understanding. For this instance of the *Error of Imprecise Expectations*, we hold fixed a criterion for evaluating a given computational model and investigate the interpretation of success for the relevant criterion. For certain benchmarks targeting natural language understanding, human performance is measured to serve as a baseline for model comparison. For example, SuperGLUE (a benchmark for "general-purpose language understanding systems"; Wang et al., 2019) has a public leaderboard which lists human performance as an aggregate score of 89.8.[4] In addition to the human baseline, a naive baseline which selects the most frequent class for each component has an aggregate score of 47.1. The *Error*

[4]The leaderboard can be found at `https://super.gluebenchmark.com/leaderboard`.

of Imprecise Expectations follows from interpreting these baselines as delineating a scale measuring language understanding from no knowledge (a score of 47.1) to human-like knowledge (a score of 89.8). This scale has two natural interpretative issues: (i) interpreting increases in degrees of success, and (ii) interpreting results which exceed the scale (a score of 90, for example).

These interpretive issues are intimately related to hype surrounding the success of neural models, as we will see shortly. However, consider the BERT baseline detailed in the paper which had an aggregate score of 69.0. For RoBERTa (a similar model to BERT; see Liu et al., 2019), an aggregate score of 84.6 was reported on the SuperGLUE benchmark. The scale bounded by human performance seems to suggest that RoBERTa is about 23% better at capturing human-like knowledge than BERT. One might be lead, then, to the conclusion that RoBERTa is more human-like than BERT, despite quite similar architecture and training data shared between the models.

Moreover, consider the fact that there are six models which have aggregate scores greater than the human baseline. Such results could suggest, non-exhaustively, that models are error-free (as compared to humans who make accidental, non-systematic, mistakes in the relevant task), that models exceed the capabilities of humans (going beyond systematic short-comings of human participants like memory limitations), or that models fail to capture meaningful variation in human performance. In this last case, the mismatch between human performance on a benchmark and something like 100% performance suggest some meaningful aspect of human capacities (e.g., perhaps some of the elements of the task themselves are errorful), so model success on such examples actually points to incorrect model behavior.[5] In failing to detail

[5]Additionally, as demonstrated in McCoy et al. (2019) a neural model can arrive at seemingly correct behavior via a wrong generalization, which again suggests that results on a benchmark are inferentially underspecified.

how one should (or how one is) interpreting model results on a benchmark, any of these three interpretations is available, though, predictably, the interpretation that models exceed humans is prevalent. For example, in discussing the SuperGLUE benchmark in April 2021, Jackie Snow reports in the Wall Street Journal[6] that by January 2021 (around two years after the creation of the benchmark), computational models "had already surpassed what most humans are able to do" – a stark example of the downstream effects of the *Error of Imprecise Expectations*.

Now, we turn to the other error in reasoning prevalent in the field, the *Error of Empirical Expansion*, which arises, primarily, in inferences from empirical results to a more general capacity of a given computational model. In the construction of a benchmark from a given domain of interest, there is an implicit restriction in the set of relevant behaviors. In natural language understanding, the set of all comprehension behaviors (in humans) is restricted to a subset of behaviors like multi-sentence reading comprehension (a subtask of SuperGLUE). In the construction of the task, there are yet further restrictions. For example, multi-sentence reading comprehension is restricted to a small paragraph with a question and a set of possible answers from which the model (or a human participant) selects the correct answer. There could be still further restrictions to certain topics, or certain types of questions (e.g., those requiring "common sense" reasoning).

In evaluating a neural model on this now quite restricted task, the resultant empirical results are similarly restricted. However, in the interpretation of results, it is quite common to forget this restriction, and instead assume (some degree of) success on the whole original domain. In our ongoing example, this could be expanding model success to generic multi-sentence reading comprehension, or more broadly, natural language understanding. This error in reasoning is often evident

[6] `https://www.wsj.com/articles/why-we-need-new-benchmarks-for-ai-11617634800`

in articles in the media which claim that neural models are "creative" (as we saw above), produce behavior "that is indistinguishable from that of a human being" (in the words of Henry Kissinger, Eric Schmidt, and Daniel Huttenlocher discussing GPT-3 in a Wall Street Journal opinion piece), or even "sentient" (as reported by a Google engineer in The Washington Post). All three of these claims correspond to very large scale capacities of models, extending far beyond what benchmarks presently test.

Finally, as an aside, I return to the issue of training data transparency, which is implicated in both errors. For example, a lack of understanding of what is contained in neural model training data both obscures the interpretation of particular results on a benchmark (relating to the *Error of Imprecise Expectations*) by potentially leaving researchers blind to data leakage (where data in the evaluation of the model is accidentally included in the training data), and also, facilitates overly expansive claims about model abilities (relating to the *Error of Empirical Expansion*) by obscuring what constitutes truly novel model output. Work within natural language processing has found that models have a high capacity to memorize large chunks of text, substantiating these concerns about data transparency (e.g., Carlini et al., 2021; Li and Wisniewski, 2021; McCoy et al., 2021).

It is also critical to briefly mention certain additional harmful effects of not adequately investigating training data, although they largely extend beyond the scope of this dissertation. Text data are not generated in a vacuum. Language users generating content are embedded in particular social contexts. Ignoring these larger social effects on language use does not reduce the influence on neural models of biased and potentially harmful data (e.g., Gehman et al., 2020; Bender et al., 2021). Training on biased data can yield models which recapitulate these same

biases (in fact they may even amplify them; e.g., Zhao et al., 2017). Moreover, the field should be mindful of who is (implicitly) meant to address these issues. Often the burden of investigating the harmful training data and model output is primarily left to those most impacted by the harm, both because those impacted are the most aware of the relevant issues (as members of socially privileged communities are often blind to the relevant issues) and because the most impacted are those who care the most (personally, morally, etc) to attempt to redress the harm (see Derczynski et al., 2022; Jakesch et al., 2022). These issues should inform any work that is meant to develop systems interacting with humans or any work creating datasets.

The errors discussed in this section (the *Error of Imprecise Expectations* and the *Error of Empirical Expansion*) are, at least in part, facilitated by a lack of clarity about the basic underlying assumptions made in work surrounding neural models. In the follow sections, I lay out what I take to be three basic assumptions addressing: (i) the relevant notion of language, (ii) the relevant comparison(s) to humans, and (iii) the relevant world view linking results in neural models to the study of humans. Some of these threads, mainly those relating to linguistic theory, are expanded upon in the concluding chapter (Chapter 6) once a sense of the empirical landscape has been established (in Chapters 3–5).

2.2 Assumptions Concerning Language

> [Language] is a fund accumulated by the members of the community through the practice of speech, a grammatical system existing potentially in every brain, or more exactly in the brains of a group of individuals;

> for the language is never complete in any single individual, but exists perfectly only in the collectivity. (de Saussure, 1983, p.13)

This dissertation is concerned with how the study of human linguistic capacities can be informed by the study of neural models of language. As human language is the relevant object of study, the first assumption considered in this chapter attempts to fix the relevant notion of language. The practical aims of natural language processing often suggest that practitioners are assuming an approach to language which centers its communicative and social function. For example, "Dialogue and Interactive Systems" and "Discourse and Pragmatics" are submission topics at the 2022 Meeting of the Association for Computational Linguistics. Therefore, it would appear the relevant notion of what language is, in the field, lies closer to the view articulated in the above quote from de Saussure (1983) and not the notion articulated in Chomskyan generative grammar. Below, I elaborate on the communicative view of language, including some of its limitations. I then argue that, despite popular framings of natural language processing work, the individualistic conception of language often associated with Chomskyan work better characterizes the common assumptions taken in the field. I conclude by noting some limitations of such an assumption and outlining its role in this dissertation.

The view that language is an object of social and communicative origin often assumes that language is defined as a pairing of sentences with socially relevant meanings, rather than an internalized generative grammar which yields a set of structured descriptions that can then be assigned meanings (as in Chomsky, 1965). Language, then, functions as a community practice learned by a child in order to communicate certain meanings reliably to other members of the same community (for a philosophical defense of this position, see Lewis, 1975). A similar

understanding of language is espoused by usage-based linguistics (e.g., Goldberg, 1995; Bybee, 2006), which has found support in natural language processing (e.g., Pannitto and Herbelot, 2020; Madabushi et al., 2020).

This section is concerned not with specific world views (see Section 2.4 for more sustained discussion of underlying theoretical commitments), but with what type of object language *is*. Therefore, I adopt the terminology in Chomsky (1986) and use *E-language* to refer to language in the sense intended by the above discussion. This is in contrast to I-language, which centers the capacities of individual language users. Put succinctly, E-language asserts that there is an object like English which exists outside of the minds of individuals, from which a child (or a computational model) can (and may only be able to) learn an approximation. E-language can evolve and change external to the mind of an individual (as seemingly assumed in work studying the cultural evolution of language; e.g., Christiansen, 2022) and, thus, is constrained (largely) by social conventions instead of biological properties of individual minds.

E-language both inherits a number of conceptual issues pointed out in Chomskyan linguistics (for extended discussion see Chomsky, 1986; Rey, 2020) and faces additional immediate challenges to its status by the dominate approaches taken in natural language processing. Turning first to the conceptual issues, there are two immediate questions: (i) what criteria determine what constitutes an E-language, and (ii) why are there restrictions on the set of sentences which are not motivated by meaning?

For the first question, consider the case of Bosnian, Croatian, and Serbian. These three languages are largely mutually-intelligible with overlapping vocabularies and grammars (despite differences in orthography) which might suggest that they are

all the same language. In fact, this was the case under Yugoslavia. However, the collapse of Yugoslavia lead to the establishment of three languages for social and political reasons (for an interesting discussion of the establishment of these languages, and the role of linguistics, see Tollefson, 2002). Such cases motivate the adage, attributed to Max Weinreich, that "a language is a dialect with an army and navy" and problematize attempts to formalize the demarcation of E-languages (similar issues are discussed in Chapter 2 of Chomsky, 1986). If E-language is taken as the relevant object of study in natural language processing, then the issue of identifying particular E-languages (and their respective communities) will have to be addressed in order for computational models to be evaluated (e.g., what E-language will a given neural model have to use in order to have an accurate and useful dialogue system).

For the second question, suppose one *could* meaningfully distinguish E-languages from one another (i.e. the above issue is resolved), the resultant language would consist of a set of sentences and meanings. However, transformational generative grammar has shown that this set of sentences is both unbounded, and also, critically, restricted in ways that go beyond distinctions in meaning. For example, the discovery of island constraints (Ross, 1967) demonstrated that sentences like "Who did stories about horrify Keisha?" are systematically ungrammatical. The view that social conventions or meaning yield such constraints is difficult to sustain (for relevant discussion, see Rey, 2020, p. 21-27). If these restrictions follow from individual capacities, rather than facts about social conventions, then modeling language will have to include accounts of language incompatible with assumptions based on E-language (for a challenge to this claim for islands in particular, though, see Wilcox et al., 2021a).

Even if the inherited conceptual issues discussed above were overcome, there remain limitations in the current approaches in natural language processing which further challenge the status of E-language. For one, while meaning is central to the establishment of an E-language, a theory of meaning is largely, unarticulated in natural language processing (though some work within computational linguistics has developed under possible world semantics; see Rooth, 2017; Collard, 2018). With regard to the evaluation of neural models of language, the notion of meaning is left unspecified, and often mixes common sense meaning, world knowledge, and linguistic meaning which may have differing statuses in an account of language.

Moreover, the association of linguistic practices (e.g., data from corpora) with a specific linguistic community is under-explored in the field, despite its centrality to E-language. Broadly, the social aspects of language are relegated to the periphery of the field (see the discussion in Hovy and Yang, 2021). Consider the development of a neural model of Spanish. Such a model would be trained on huge amounts of unstructured text data that mingle speech from a variety of communities. Sociolinguistic work has shown that rates of pro-drop differ quite substantially across communities (e.g., Otheguy et al., 2008; Mayol, 2012). How then should we expect this neural model to behave with respect to this grammatical phenomenon? One option is for the neural model to develop a rate of pro drop that tracks the average across these communities. Alternatively, we might want the model to develop behavior conditioned on specific dialectics. The creation of training data and the development of models does not specify a commitment in either direction, thus, the field is missing a constitutive aspect of the study of E-language.

This section should not be taken as discouraging an approach to natural language processing which is concerned with E-language. In fact, explicitly taking E-language

as the object of study would beneficially center the issues of community and language use. Failing to account for these aspects of language in building models can exclude large portions of the population of language users from being able to benefit from novel technology. Automatic speech recognition is an instructive case in point. Standard American English is commonly assumed in the development of these technologies, which can cause downstream usage issues for speakers of non-standard Englishes, like African American Vernacular English (see Koenecke et al., 2020). Rather, the above discussion aims to demonstrate that the work required for this type of approach (constructing socially annotated training data, building models which account for the practices of a variety of communities, etc.) has not yet been widely undertaken. Thus, despite a close connection with E-language in the overt framing of the field, most work does not commit to taking E-language as its object of study (or ignores the preliminary steps necessary in the development of such an approach). I will instead argue that I-language (the capacities of an individual language user) better characterizes the approach taken by the field.

I-language, as presented in Chomsky (1986), denotes the concept that language is internal (to the mind), (possessed by an) individual, and intensional (i.e. a generative capacity). Assuming I-language places the grammar of individual humans as the object of study, rather than the practice of a set of people as discussed above. Thus, in evaluating a neural model under this conception of language, a given computational model would be compared to the ability of an individual language user to produce and interpret sentences. While not often explicitly discussed as such, this is exactly the framing commonly assumed in work interpreting neural models for linguistic knowledge. For example, a number of studies have investigated the relationship between the behaviors and representations in neural models and those observed from individual humans (e.g., Schrimpf et al., 2020; Wilcox et al.,

2020a; Heilbron et al., 2020). Additionally, in benchmark creation (as discussed in Section 2.1) human performance is often taken as a baseline (e.g., Wang et al., 2018, 2019). In fact, even laypeople seem to understand neural models as possessing something like I-language, and thus, assume that successful models have the same capacities as individual humans (e.g., neural models are often said to "read" or "write"). I therefore assume in this dissertation that I-language is the relevant object of study.

It is also important to note that using I-language as the object of study is not without problems. Implicit in the comparisons between neural models and individual humans is the assumption that the training data for models contain the relevant signals necessary to uncover an I-language. This link is largely left unexplored in the field. Rather, it is taken for granted that training data contain the relevant information (at least to arrive at linguistic form; see Bender and Koller, 2020), and model behavior reflects the capacity for a neural model to extract this information.[7]

In reality, training data actually contain a mix of I-languages (a similar issue to training data containing many E-languages discussed above). The extent this should concern us may differ by the status of the relevant linguistic process (i.e. whether the process under investigation is a property of the dominant I-language represented by the training data). For some processes, like subject-verb agreement in English, this seems rather innocuous. However, even for subject-verb agreement this assumption can be problematic. There are I-languages which we group under the E-language English which do not have the same verbal agreement, so by assuming this behavior is desirable we are, in fact, enforcing a normative judgment about English. The relationships between data, I-language, and variation are complex and cannot be

[7]I return to the relationship between training data and model behavior in Chapter 6

resolved in this dissertation. However, it would seem to me that any answer must begin by investigating training data more carefully (as is advocated throughout this work).

In sum, I assume that I-language is the most viable conception of language for this dissertation. Crucially, I am not claiming neural models have an I-language; rather, human I-languages are the relevant point of comparison. As will be seen, there are mismatches between neural models and humans that cast doubt on the ability of neural models to acquire a human-like I-language. In the following section, I discuss more directly the relevant comparisons between neural models and humans assumed in this dissertation.

2.3 Assumptions Concerning Comparisons to Human Linguistic Processing

> If I write "I keep a giraffe in my pocket," you are able to understand me despite the fact that, on even the most inflationary construal of the notion of context, there is nothing in the context of inscription that would have enabled you to predict either its form or its content. (Fodor, 1983, p. 67)

While this dissertation lays out concrete empirical instances of mismatch between neural models and humans, we may ask whether and how we can characterize the linguistic structures or behaviors that escape the capacity of neural models of language. I assume that such a characterization requires at least three components: (i) the computational model and its "inductive biases" (e.g., McCoy et al., 2020),

(ii) the nature of linguistic data (see Chapter 6 for further discussion), and (iii) the desired level of human-like linguistic processing. It is this last component which motivates the second question at the start of this chapter – what types of human linguistic processing are relevant in evaluating neural models. In what follows, I discuss the use of neural models as cognitive models of humans (though this dissertation does not assume this position) and the role of Surprisal Theory (Hale, 2001) in mediating comparisons between neural models and humans. Then, I frame the comparisons relevant to this dissertation by addressing the levels of linguistic representation implicated in human linguistic processing.

There are two main reasons for comparing neural models to humans. The first, uses a measurement of human behavior as a baseline to evaluate model performance, addressing something like whether a model is accurate. An example of this can be found in the discussion of natural language understanding benchmarks in Section 2.1. The second conceives of neural models as *cognitive models* of humans. While this dissertation does not assume that neural models are a model of human cognition, framing of neural models along those lines is often assumed in work evaluating models with human behavioral measures (e.g., reading times, fMRI signal).

Early work on the use of neural models as cognitive models proceeded by creating (often by hand) connectionist models which accounted for aspects of human behavior (e.g., Parallel Distributed Processing Rumelhart and McClelland, 1986). In modeling language, in particular, seminal work in Elman (1990) trained recurrent neural models on unstructured sentences to demonstrate that aspects of linguistic knowledge could follow from emergent properties of data. Recent advances in natural language processing have yielded much larger and more capable models, which has driven similar claims about human linguistic knowledge following

from data (e.g., Wilcox et al., 2019a, 2021a).

The current approach in interpreting neural models (within natural language processing) can be characterized, then, as reductive, that is, aiming to reduce the "complexity" of linguistic theory by showing naive models trained on unstructured data can yield the same linguistic behaviors as humans. However, there is nothing in the approach generally which requires this reductive tendency. Neural models, themselves, can be built to encode hierarchical structure which allows for a richer system than what is commonly desired by practitioners (e.g., Dyer et al., 2016; Kim et al., 2019). This dissertation follows the bulk of interpretability work in using neural models without any explicit and predefined linguistic structure. I return to these matters in the follow section.

Assuming we want to make comparisons between humans and neural models (for either of the two reasons mentioned above), we need some way of relating the output of neural models which is, typically, a probability distribution over words given a context, to behavioral measures of humans. One common linkage is facilitated by Surprisal Theory (e.g., Hale, 2001; Levy, 2008). Surprisal Theory relates cognitive effort (measured via reading times, EEG signal, etc) to the reduction in possible alternative parses (or other linguistic representations) necessitated by the incorporation of unfolding language.

Consider, the classic garden path sentence "the horse raced past the barn fell" (from Bever, 1970). English speakers reading this sentence incrementally (i.e. word by word) experience increased cognitive effort at the final word "fell" as evidenced by increased reading times. We can explain this increased cognitive effort by noting that before "fell" there are two alternative parses for the sentence: one which labels "raced" as a main verb, and another which places "raced" within a reduced relative

clause (as in "the horse that was raced past the barn"). Following the discussion in Hale (2001), Surprisal Theory associates the cognitive effort at "fell" with a reduction in the two alternative parses to the one parse which has "raced" as part of a reduced relative clause. This effort is further modulated by the likelihood of these alternative parses. The main verb reading is more probable than the reduced relative clause reading, so this induces a greater change in the cognitive system. Surprisal Theory, then, relates the probability of a word (conditioned on the preceding context) with a human behavioral measure. It is this connection which allows one to take the probabilities outputted by a neural model and compare them to human linguistic behavior.

With a general linkage between neural model outputs and human linguistic predictions established via Surprisal Theory, we may now ask which specific behaviors we should use in evaluating models. In characterizing human linguistic behaviors in psycholinguistics, it is common to refer to levels of linguistic representation. The relevant levels include syntax, semantics, and discourse (which may include both larger contextual effects on processing, but also, socio-indexical information). The relationship between linguistic processing and these levels of linguistic representation remains a contentious area of research.

For example, a number of theoretical accounts have attempted to clarify at what time in linguistic processing each level exerts an influence. Often syntax is taken as the primary level of representation implicated in parsing behavior (e.g., the use of probabilistic context free grammars to account for garden path phenomenon in Hale, 2001). However, the proposals range from models which assume syntactic structure is built without consideration of semantic or pragmatic plausibility (e.g., Frazier and Fodor, 1978) to models where all linguistic levels operate in parallel as

constraints (e.g., McClelland et al., 1989).

Empirical work suggests that semantic information can operate rapidly to constrain predictions about upcoming linguistic material (e.g., Altmann and Steedman, 1988), so regardless of the relative importance of each level, very early components of linguistic processing can reference multiple levels of linguistic representation. Moreover, experiments suggest that even contradictory linguistic information can be sustained in human linguistic processing. Consider the effect of local coherence demonstrated in Tabor et al. (2004) for sentences like "the coach smiled at the player tossed a frisbee by the opposing team". This sentence contains a locally coherent active clause interpretation that"the player tossed a frisbee" which is not licensed by the whole string (where the intended meaning is "the coach smiled at the player who was tossed a frisbee by the opposing team"). Human readers can be distracted by these locally coherent yet globally impossible parses. Incremental processing in humans, then, is constituted by a range of (sometimes contradictory) linguistic information.

Thus, in evaluating neural models with reference to humans, care must be taken to tease apart which linguistic representations are implicated by the relevant comparison. Presently, both immediate linguistic processing, via comparison to incremental reading times, and later linguistic behavior, via comparison to linguistic judgments, can be taken as relevant targets.[8] In fact, the same underlying computational model has been used in recent work to account for both grammaticality judgments (e.g., Warstadt et al., 2020a) and human behavioral measures (e.g., Wilcox et al., 2020a). Linguistic behaviors even more downstream of immediate processing considerations like judgments of whether a sentence entails another or

[8]For interesting discussion of the relationship between linguistic judgments and linguistic theory see Chapter 3 of Ludlow (2011)

whether a given sentence is an adequate paraphrase of a paragraph have been investigated (e.g., the case of natural language understanding discussed in Section 2.1).

In the following chapters, the interpretation of results from neural models is facilitated by explicitly identifying the relevant level (and time course) of linguistic processing under investigation. As a basic hypothesis, I expect that only certain aspects of human linguistic processing will be evidenced by neural models. That is, I do not consider them models of all of human cognition. Thus, there are presumably certain behaviors which require additional modeling components. The form this limitation takes depends on the empirical landscape, but possibilities are that only certain types of relations (attachment or association, following Construal Theory Frazier and Clifton, 1996) or only immediate syntactic information (to the exclusion of pragmatic or semantic information, following a modularity thesis Fodor, 1983) are captured by current neural models. In this dissertation, only linguistic processes evidenced by earlier aspects of linguistic processing (i.e. those not requiring additional reasoning by a human reader) are considered, leaving to future work additional aspects of human linguistic knowledge.

2.4 Assumptions Concerning Inferences About Human Capacities

> [I]n the study of language, we cannot aspire to 'explain' the presence and structure of language as a composite function of various descriptively isolable language behaviors. (Bever, 1970, p. 280)

In this section, I return to the final question posed in this chapter: what world view mediates inferences from computational models to human capacities? Put another way, if, in studying neural models, we aim to learn something about humans, it is necessary to lay out the world view which facilitates this connection. In the field, neural models are often taken as explaining (aspects) of how humans come to acquire their linguistic knowledge (see, for example, the discussion in Wilcox et al., 2021a and Linzen and Baroni, 2021).

In characterizing the origin of human linguistic knowledge, there are two main world views. The first, often called nativism and associated with Chomskyan linguistics, attributes to the human mind/brain an innate capacity for acquiring a linguistic system, which may include concepts like parameters or certain language specific mechanisms (e.g., Chomsky, 1965, 1995, 2000). The other world view, attributes very little (if any) innate capacities or language specific mechanisms to humans (for an example of such a view see Goldberg, 2003). I follow Rey (2020) in referring to this world view as *superficialism*. Superficialism asserts that "all genuine psychological distinctions can be made on the basis of ordinary behavior or introspection" (Rey, 2020, p. 93). It is a successor of sorts to Behaviorism, which fell out of favor due in part to arguments in Chomsky (1959).

I distinguish between nativism and superficialism in this section, primarily, by the role empirical data plays in their accounts of language acquisition. Nativism, while noting that primary linguistic data certainly influences a child's grammar (I know English and not Portuguese after all), asserts that the innate capacities of individuals largely determine the acquisition of language. Conversely, superficialism asserts that language knowledge is primarily constituted by aspects of data and not our biology.

Within natural language processing, the nature of most neural models clearly conforms to the world view of superficialism. The neural models investigated in this dissertation, for example, have no pre-defined concept of hierarchical linguistic structure, and training data is presented as a linear sequence of words (i.e. as a sentence) rather than in something like the form of a syntactic tree. In effect, the world view assumed in the study of neural models is a quite extreme version of the superficialism discussed in Rey (2020). Linguistic structure is assumed to be evidenced entirely on the basis of surface contrasts in language use with no additional room for something like "introspection" in the model architecture or organizing principle guiding the formation of a linguistic system (e.g., the linguistic system of neural models has no explicit requirement for simplicity).

In adopting superficialism, one is immediately faced with challenges, both from linguistic theory and from empirical results in the study of neural networks. This dissertation, itself, provides empirical challenges to this world view. Below, I briefly discuss some theoretical and practical challenges, before returning to the role of superficialism in this dissertation and beyond.

The limitations to a world view like superficialism have long been noted in Chomsky's work. One salient way this is done is via discrepancies between linguistic data and the linguistically meaningful generalizations speakers make (i.e. the generalizations made in the acquisition of an I-language). Consider the following examples from Chomsky (1986):

(1) a. John ate an apple

b. John ate

c. John is too stubborn to talk to Bill

d. John is too stubborn to talk to

In (1), consider the disparate role of deletion. A noun phrase object (*an apple*) is missing between (1-a) and (1-b). The meaning of (1-b) seems to relate to the meaning of (1-a), where instead of eating a specific object, John is taken to have eaten a generic object. Assuming that deletion of an object results in the interpretation of some unspecified, arbitrary object that could be an object of the verb (i.e. a generalization of what we see in going from (1-a) to (1-b)), (1-d) should mean John is too stubborn to talk to any arbitrary person. However, (1-d) means that John is too stubborn for some arbitrary person to talk to him (John). In other words, surface relatedness does not necessarily suggest structural relatedness.

In addition to mismatches in form, as in (1), there are certain constructions which seem to have no evidence in surface contrasts. For example, the Binding Principles (Chomsky, 1981) are restrictions on certain meanings, rather than on certain surface forms. That is, it is perfectly fine to produce the sentence *John likes him*; that sentence just cannot be taken to mean that *John likes John*.

I turn now to the challenges to superficialism posed by work in natural language processing. The debates surrounding whether neural models acquire human-like linguistic meaning are instructive in this regard. Neural models have no concept of intentional action, and text data alone seems to provide no relevant clues. Yet intention is core to understanding the communicative meaning and use of language (Bender and Koller, 2020). Moreover, aspects of language which are informed by our bodies (e.g., colors) and by our place in social relations (e.g., the contexts where polite speech is necessitated) are also missing in the text data commonly used to train data (Bisk et al., 2020). Even in ignoring the issue of

meaning, non-linguistic generalizations salient in training data can prevent the use of linguistically meaningful properties (e.g., McCoy et al., 2019) or delay the acquisition of linguistic generalizations (e.g., Warstadt et al., 2020b). The following chapters in this dissertation provide yet more issues in using linguistic data alone as the source of linguistic knowledge.

Despite these limitations, superficialism remains the most viable world view associated with neural models in natural language processing. This dissertation is meant to demonstrate that explicitly relating the study of neural models to superficialism is, in fact, beneficial for researchers. It both allows for results from neural models to be understood as advancing a particular world view and allows one to test the viability of superficialism more broadly. Interestingly, much of the work relating neural models to claims about human linguistic knowledge conceive of results from neural models as, instead, challenging nativism. In fact, it seems more appropriate to think of these results as providing evidence for superficialism, as nativism is not explicitly tested in current work, nor can it be given the nature of the models and the data mentioned earlier in this section (e.g., neural models which instantiate Minimalist Syntax are not developed and tested for their coverage of corpus data).

Ultimately, this dissertation argues that superficialism fails to capture meaningful aspects of human linguistic knowledge, and because this view is the only viable world view associated with neural models, such models are inherently limited in their ability to acquire human-like linguistic systems. Moreover, it appears that linguistic data provides evidence for linguistic systems that humans entirely ignore pointing to additional misalignment between human linguistic knowledge and the information available in data.[9]

[9]I return to this discussion in Chapter 6 in relating the results of this dissertation to linguistic

2.5 Summary

In the following chapters, a number of assumptions are necessary in order to fully interpret the results of neural models (and their relationship to human linguistic knowledge). For one, I take the underlying object of study to be the linguistic system internalized by individual speakers of a language (i.e. I-language) as opposed to the study of community practices and situations of language use (i.e. E-language). Additionally, I compare neural models to measures of individual linguistic processing, taking note of the explicit time course (or linguistic representations) implicated in the relevant comparison, rather than make comparisons to inferences speakers gather from occurrences of specific utterances. Finally, in order to infer something about human capacities from investigating neural models, I assume that superficialism must be the relevant world view.

Certainly, these assumptions are not the only way one may understand neural models. Nonetheless, it is the approach advanced in this dissertation. Ultimately, human linguistic systems are complex, following both from human-specific (and internal) mechanisms and representations, and from the contents and character of experience. Therefore, it is not surprising that neural models fall short of human-like linguistic knowledge. These failures are interesting insofar as they illuminate properties of data, and therefore experience, which appear to be systematic. The nature of this systematicity calls out for some principled explanation in our theories and in our computational models.

---

theory.

CHAPTER 3

IMPLICIT CAUSALITY

3.1 Introduction

Work on probing large pre-trained models for linguistic knowledge (e.g., Gulordava et al., 2018; Futrell et al., 2018b; Hu et al., 2020b) has focused on isolated linguistic phenomena that are constructed to target a single linguistic process or representation (as is common in psycholinguistics).[1] For example, BLiMP (a popular evaluation dataset for probing the linguistic knowledge of neural models; Warstadt et al., 2020a) sets up minimal pairs for anaphoric agreement by manipulating the gender and number of anaphora (cf. *Many girls insulted* ***themselves****.* vs **Many girls insulted* ***herself****.*). Comparing stimuli which only differ in the target process (e.g., the probability of the sentences ending with grammatical or ungrammatical anaphors) is meant to provide evidence of models' linguistic knowledge. When models assign more probability to the grammatical sentence it is because they know, hopefully non-trivial, aspects of the relevant linguistic process, and when they fail to distinguish grammatical and ungrammatical sentences it is because they do not know the process.

Consider a linguistic process that is evidenced by more than one (surface) linguistic behavior. Suppose further that, for a given neural model, there is overlap between humans and neural models for only one of the surface relevant contrasts. Given the paradigm above (of checking isolated behaviors), we have no way of

[1] Code for replicating the experiments, figures, and statistical models in this chapter can be found on Github at `https://github.com/forrestdavis/Dissertation/tree/main/ImplicitCausality`. Additional materials are provided in Appendix A. Parts of this chapter appear in two published works: Davis and van Schijndel (2020a) and Davis and van Schijndel (2021).

understanding model behavior when faced with paradoxical results. Moreover, in observed utterances it is rare to find that only one process, alone, influences surface behavior. Subject-verb agreement, for example, requires something like a number distinction applied (correctly) to nominal categories in conjugation with hierarchical structural representations to pick out the verbal agreement controller. The core problem is that linguistic processes are interactive, in the sense that they are interleaved and co-constituted. This interaction can be obscured by narrow investigations of isolated linguistic processes.

This chapter attempts to remedy this shortcoming by focusing on interaction. We explore how a single discourse structure interacts with other linguistic processes, both in English and cross-linguistically. The particular discourse structure is governed by implicit causality (IC) verbs (Garvey and Caramazza, 1974). Such verbs influence pronoun production and comprehension:

(1) a. Sally frightened Mary because she was so terrifying.
    b. Sally feared Mary because she was so terrifying.

In (1), *she* agrees in gender with both *Sally* and *Mary*, so both are possible antecedents. However, English speakers overwhelmingly interpret *she* as referring to *Sally* in (1-a) and *Mary* in (1-b), despite the semantic overlap between the verbs. Verbs that have a subject preference (e.g., *frightened*) are called subject-biased IC verbs, and verbs with a object preference (e.g., *feared*) are called object-biased IC verbs.

This chapter begins by probing neural models of English for knowledge of IC. We find that transformer models show some degree of IC knowledge. We then

explore how IC interacts with syntax in English, before broadening our investigation to Chinese, Italian, and Spanish. While transformer models of English are able to capture the influence of IC verb bias on pronoun prediction (including the gradience in human IC verb biases) similar neural models of languages other than English fail to capture IC and its interactions.[2] Ultimately, this chapter provides evidence for a mismatch between acquiring an isolated linguistic process and acquiring a linguistic system (of sometimes competing linguistic processes).

3.2 Background

In evaluating neural models for an implicit causality bias, it is natural to ask whether neural models can be reasonably expected to learn IC at all. We begin by addressing the psycholinguistic evidence for IC with an emphasis on the cross-linguistic distribution of IC and the relationship between IC and the linguistic system as a whole. As IC lies at the intersection of syntactic and coreference processing, we then turn to the evidence within computational linguistics and natural language processing that syntactic and coreferential phenomenon can be learned by neural networks. Finally, we address the effect competing linguistic (and non-linguistic) processes have on neural models, which motivates the later experiments in this chapter. Ultimately, the literature within psycholinguistics and computational linguistics suggests that learning IC should be possible (at least in principal) for neural models.

Implicit causality is a well established phenomenon in the psycholinguistic

[2]Despite the overlap between humans and neural models of English in this chapter, IC knowledge is again not robust. In Section 4.6, we demonstrate that the interaction between IC and relative clause attachment in English is not learned by neural models.

literature (e.g., Garvey and Caramazza, 1974; Kehler et al., 2007; Ferstl et al., 2011; Hartshorne and Snedeker, 2013; Hartshorne, 2014; Williams, 2020). Within this literature, IC has been shown to be remarkably consistent cross-linguistically (see Hartshorne et al., 2013; Ngo and Kaiser, 2020). That is, IC verbs (with similar levels of bias) have been attested in a variety of languages, including Korean (Yi and Koenig, 2020), Japanese (Hartshorne et al., 2013), Vietnamese (Ngo and Kaiser, 2020), and American Sign Language (Frederiksen and Mayberry, 2021). In this chapter, we investigated Italian (Mannetti and De Grada, 1991), Spanish (Goikoetxea et al., 2008), Chinese (Hartshorne et al., 2013), and English (e.g., Garvey and Caramazza, 1974).

Current accounts of IC in psycholinguistics claim that the phenomenon is inherently a linguistic process, which does not rely on additional pragmatic inferences by comprehenders (e.g., Rohde et al., 2011; Hartshorne and Snedeker, 2013). Thus, IC is argued to be contained within the linguistic signal, analogous to evidence of syntactic agreement and verb-argument structure within corpora. We hypothesize that if these claims are correct, then neural models will be able to behave in accordance with the IC bias documented in human psycholinguistic studies. Moreover, given the cross-linguistic consistency of IC for humans, we expect neural models of a variety of languages to demonstrate a consistent IC bias. However, we find that only models of English and Chinese, and not those of Italian and Spanish have a human-like IC bias.

While we established above that IC verb biases are both well documented in the psycholinguistic literature and tied closely to the linguistic signal, we have not yet addressed the existing support for positing that neural models should be able to learn IC. IC lies at the intersection of coreferential and syntactic processing.

Aspects of both of these domains have been claimed to be encoded in neural models. The ability of neural models to encode coreferential knowledge has largely been explored in the domain of coreference resolution. Prior work has suggested that neural models can learn coreference resolution to some extent (e.g., Peters et al., 2018; Sorodoc et al., 2020). In the present study, we focus on within-sentence resolution rather than the ability of neural models to track entities over larger spans of text (cf. Sorodoc et al., 2020). At this granularity of coreference resolution, LSTM neural models strongly favor reference to male entities (Jumelet et al., 2019), for which the present study finds additional support. In this chapter, rather than utilizing a more limited modeling objective such as coreference resolution (cf. Cheng and Erk, 2020), we followed Sorodoc et al. (2020) in focusing on the referential knowledge of models trained with a general language modeling objective.

With regards to syntactic processing and representations, a growing body of literature suggests that neural models are able to acquire syntactic knowledge. In particular, subject-verb agreement has been explored extensively (e.g., Linzen et al., 2016; Bernardy and Lappin, 2017; Enguehard et al., 2017) with results at human level performance in some cases (Gulordava et al., 2018). Additionally, work has shown human-like behavior when processing reflexive pronouns, negative polarity items (Futrell et al., 2018b), center embedding and syntactic islands (Wilcox et al., 2019b,a). This literature generally suggests that neural models encode some type of abstract syntactic representation (e.g., Prasad et al., 2019). Additionally, recent work has shown neural models learn linguistic representations beyond syntax, such as pragmatics and discourse structure (Jeretic et al., 2020; Schuster et al., 2020; Davis and van Schijndel, 2020b).

Finally, we turn to competition in linguistic and non-linguistic processes within

neural models. Existing work has shown that non-linguistic biases of neural models mimic English-like linguistic structure, limiting the generalizability of claims founded on English data (e.g., Dyer et al., 2019; Davis and van Schijndel, 2020c). Thus, claims of the success of certain models (or architectures) can be strengthened by looking beyond English, where structures may have differing surface manifestations. Moreover, competition between non-linguistic and linguistic generalizations has been documented (e.g., spurious correlations internal to the creation of an evaluation dataset interacts with linguistic generalizations like entailment; e.g., McCoy et al., 2019; Davis and van Schijndel, 2020a; Warstadt et al., 2020b). The findings in Warstadt et al. (2020b), that linguistic knowledge is represented within a model much earlier than attestation in model behavior, bears resemblance to our claims in this chapter. We find that linguistic knowledge can, in fact, lie dormant due to other linguistic processes (and not just non-linguistic biases) in a language. It would appear, then, that some linguistic knowledge may never surface in model behavior even with the presence of corresponding representations, though further work is needed on this point.

Finally, in the construction of our experiments, we were inspired by work with synthetic language data meant to evaluate the underlying linguistic capabilities of language models (e.g., McCoy et al., 2018; Ravfogel et al., 2019). We made use of modified versions of languages that accentuated, or weakened, evidence for certain linguistic processes. The goal of such modification in our work is quite similar both to work which attempts to remove targeted linguistic knowledge in model representations (e.g., Ravfogel et al., 2020; Elazar et al., 2021) and to work which investigates the representational space of models via priming (Prasad et al., 2019; Misra et al., 2020). The present chapter differs in noting how linguistic representations interact to influence model behavior, rather than in noting how

linguistic representation are encoded in models.

3.3 IC Behavior of Neural Models of English

Current work in psycholinguistics on IC verbs suggests that IC bias is analogous to verb argument structure, so we hypothesized that a human-like IC bias should be learnable from text data alone (see Hartshorne and Snedeker, 2013; Williams, 2020). In what follows, we focused on English and investigated both autoregressive neural models (e.g., GPT-2, LSTMS) and non-autoregressive (e.g., BERT) transformer neural models. We break the results into a categorical and a gradient investigation of IC behavior. The categorical investigations sought broad evidence of an effect of subject or object biased IC. The gradient investigations addressed the overlap between neural model predictions and the gradient IC biases found in Ferstl et al. (2011). To look ahead, we found that transformers models behave, at least to some degree, like humans in pronoun prediction following IC verbs, while LSTM neural models fall short, showing no effects of IC bias. Moreover, transformer neural models capture both categorical and gradient aspects of IC verb biases.

3.3.1 Methods: Neural Models of Language

We trained 25 LSTM language models on the Wikitext-103 corpus (Merity et al., 2016) with a vocabulary constrained to the most frequent 50K words.[3] Additionally, we used two pre-trained autoregressive transformer neural models: TransformerXL

[3]The models had two LSTM layers with 400 hidden units each, 400-dimensional word embeddings, a dropout rate of 0.2 and batch size 20, and were trained for 40 epochs (with early stopping) using PyTorch. The mean perplexity for the models on the validation data was 40.6 with a standard deviation of 2.05.

(Dai et al., 2019) and GPT-2 XL (Radford et al., 2019). We also used two popular non-autoregressive transformer model variants, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). All transformer models were accessed via HuggingFace (Wolf et al., 2020).

TransformerXL was trained on Wikitext-103, like our LSTM models, but had more parameters and a larger vocabulary. GPT-2 XL differs from the other models in lacking recurrence (instead utilizing non-recurrent self-attention) and in amount and diversity of training data (1 billion tokens compared to the 103 million in Wikitext-103). BERT and RoBERTa differ from all the other models in their treatment of context, conditioning on both the left and right contexts. Like GPT-2 XL, BERT and RoBERTa are trained on a large amount of text (3.3 billion and 30 billion tokens respectively). As such, we caution against extracting explicit, mechanistic claims from this chapter concerning the relationship between learned linguistic knowledge and model configurations and training data. Instead, this section points to apparent differences between transformers and LSTMs with regard to use and acquisition of discourse structure, leaving explanatory principles to further work.

3.3.2 Methods: Stimuli

Our data consisted of the stimuli from a human experiment conducted in Ferstl et al. (2011), which asked participants to give continuations of sentence fragments of the following form:

(2) Kate accused Bill because ...

Continuations were coded across 305 verbs for whether participants referenced the subject (i.e. *she*) or the object (i.e. *he*).[4] The results of this coding were then converted into a bias score for each verb, ranging from 100 for verbs whose valid continuations uniquely refer to the subject (i.e. subject-biased) to -100 for verbs whose valid continuations uniquely refer to the object (i.e. object-biased).

In the present study, we took 246 of these verbs[5] and generated stimuli as in (2) using 14 pairs of stereotypical male and female nouns (e.g., *man* vs. *woman*, *king* vs. *queen*), rather than rely on proper names as was done in Ferstl et al. (2011).[6] We created two categories of stimuli: those for autoregressive models (e.g., LSTM) and those for non-autoregressive models (e.g., BERT) all with differing gender[7] resulting in 6888 sentences per category. For autoregressive models we truncated the sentences at the pronoun (e.g., *the woman accused the man because she*). For non-autoregressive models, we used a neutral right context, *was there.*[8] An example stimuli for non-autoregressive models was:

(3) the man admired the woman because [MASK] was there.[9]

[4] An additional category, other, was included for ambiguous (e.g., *they hate each other*) or non-referential continuations (e.g., *it was a rough day*).

[5] 59 verbs were outside of our LSTM model vocabulary, so they were excluded.

[6] See Appendix A.1 for all the noun pairs and verbs.

[7] We balanced our stimuli by gender, so we had the same number of female subjects as male subjects and similarly for objects.

[8] Using *here, outside,* or *inside* as the right context produces qualitatively the same patterns.

[9] The model specific mask token was used. Additionally, all models were uncased, with the exception of RoBERTa, so lower cased stimuli were used.

3.3.3 Methods: Measures

We evaluated the models' external behavior (e.g., predicted next-words) by comparing the probability of *he* and *she*. For autoregressive models, the probability is conditioned on the preceding context:

$$P(\text{pronoun}) = P(w_i = \text{pronoun} | w_1 \cdots w_{i-1}) \tag{3.1}$$

For non-autoregressive models, we gathered the probability assigned to *he* and *she* at the MASK location, which is conditioned on both the left context (containing the IC verb) and the right context (e.g., *was there*; see Section 3.3.2).

For our categorical experiments (see Section 3.3.4), we focused on the IC effect on pronouns by calculating the probability of pronouns referring to the subject and pronouns referring to the object for each IC verb. We predicted that IC verbs would influence the probability of pronouns, with subject-biased IC verbs increasing the probability of pronouns agreeing with the subject, and object-biased IC verbs increasing the probability of pronouns agreeing with the object. For example, BERT assigned to (5) a score of 0.01 for the subject antecedent (i.e. *he*) and 0.97 for the object (i.e. *she*), in line with the object-bias of *admire*. This methodology follows subject-verb agreement experiments, where verbal agreement is investigated in contrastive conditions (e.g., *Cats are* vs. **Cats is*; Linzen et al., 2016; Mueller et al., 2020)

For our gradient experiments (see Section 3.3.5), we compared the strength of each verb's IC bias for neural models to the gradient IC bias reported in the human experiments. We took the difference between the pronoun referring to the subject and the pronoun referring to the object for each verb and multiplied the

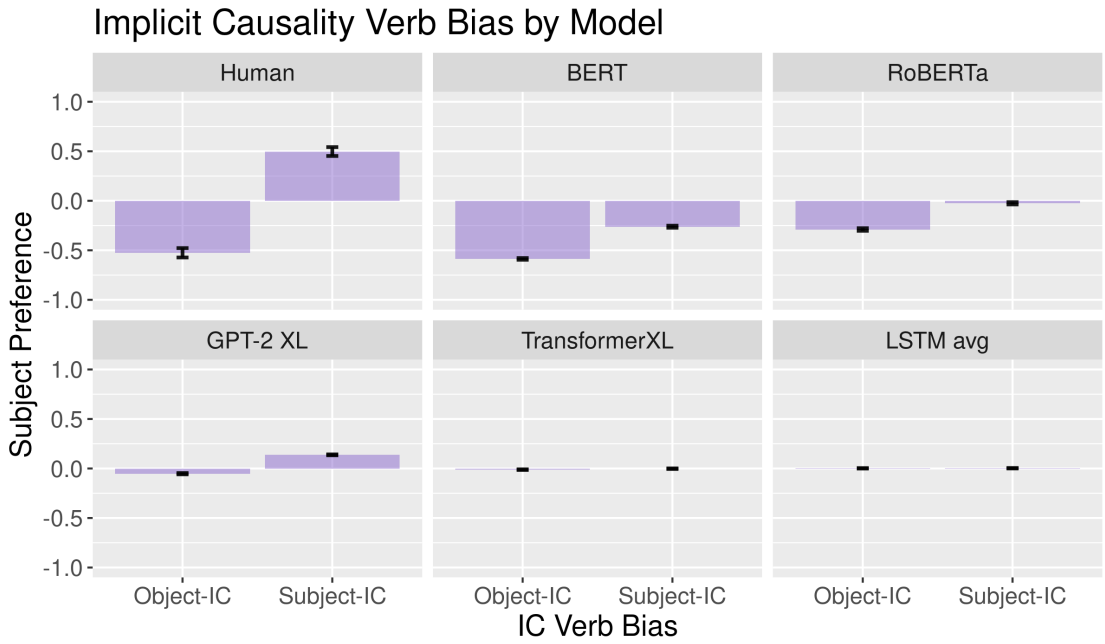

Figure 3.1: Subject preference grouped by implicit causality verb type for humans (from Ferstl et al., 2011), BERT, RoBERTa, GPT-2 XL, TransformerXL, and the by-item average of LSTMs. A value of 1.0 corresponds to a complete preference for pronouns agreeing in gender with the subject (i.e. a subject bias), and a value of -1.0 corresponds to a complete preference for pronouns agreeing with the object (i.e. an object bias). Error bars are 95% confidence intervals.

value by 100.[10] This measure ranges from 100 for verbs which fully prefer pronouns that refer to subjects (i.e. a subject-biased IC verb) and -100 for verbs which fully prefer pronouns that refer to objects (i.e. a object-biased IC verb), following the human measure which assigns 100 to fully subject-biased IC verbs and -100 to fully object-biased IC verbs. For the example of *admire*, BERT has a value of -96 and the human measure from Ferstl et al. (2011) is -92, suggesting a similar gradient bias.

[10]The same verb occurred with different subject and objects, so the final value is the average of the differences between pronouns referring to the subject and pronouns referring to the object for the same subject, verb, and object.

3.3.4 Categorical Influence of IC on Model Behavior

While the source experimental investigation for this section reported gradient IC bias scores (ranging from 100 to -100; see 3.3.2), we began by looking for categorical effects of subject or object IC bias in the prediction of unambiguous pronouns. If models conform to the IC bias behavior of humans, then pronouns agreeing with the object should be more likely after object-biased IC verbs and pronouns agreeing with the subject should be more likely after subject-biased IC verbs.

Results for each neural model (BERT, RoBERTa, LSTMs, TransformerXL, GPT-2 XL; see Section 3.3.1) are given in Figure 3.1. Statistical analyses[11] were conducted via linear-mixed effects models.[12] TransformerXL, the LSTMs, and GPT-2 XL all had main effects of gender, with a preference for masculine pronouns in TransformerXL and the LSTMs and a preference for feminine pronouns in GPT-2 XL. This corroborated existing claims in the literature (see Jumelet et al., 2019). As is visually apparent in Figure 3.1, there is variation in the IC conditioned behavior of the various neural models investigated.

The LSTM language models showed no significant interaction effect between IC verb bias and antecedent, or any main effects of IC verb bias and antecedent. This suggests that the LSTM language models failed to learn any human-like IC bias. TransformerXL did show a significant interaction effect between IC verb bias and antecedents where subject-biased IC verbs increased the probability of subject pronouns as compared to object-biased IC verbs. There was, however, no effect for

[11] We used lme4 (version 1.1.23; Bates et al., 2015) and lmerTest (version 3.1.2; Kuznetsova et al., 2017) in R.

[12] We fit a model to predict the probability of the pronoun with a three-way interaction between IC type (subject or object), pronoun antecedent (subject or object), and gender (male or female) and with random slopes for verbs and noun pairs (e.g., man and woman). For the LSTMs random slopes for verbs, noun pairs, and model were included. Post-hoc t-tests were conducted to evaluate effects. The threshold for significance was set at 0.005.

object-biased IC verbs and object pronouns, suggesting that the local agreement in TransformerXL is stronger than any IC-bias.

For the other transformer models, there were statistically significant interactions between IC verb bias and antecedent. GPT-2 XL fully matched the qualitative human pattern with pronouns referring to the object being preferred to those referring to the subject after object-biased IC verbs. Similarly, pronouns referring to the subject were more likely than those referring to the object after subject-biased IC verbs. BERT and RoBERTa both had a recency bias, generally preferring pronouns that refer to objects. However, pronouns referring to the subject were more likely after subject-biased IC verbs than object-biased IC verbs. Additionally, pronouns referring to the object were more likely after object-biased IC verbs than subject-biased IC verbs.

3.3.5 Gradient Influence of IC on Model Behavior

As discussed above in Section 3.3.3, we gathered gradient IC bias scores for each IC verb in order to compare models to the gradient IC biases reported in Ferstl et al. (2011). If models capture the gradient strength of IC bias for each verb, then we should find significant correlations between the model IC bias and the human IC biases.

Results for each neural model (BERT, RoBERTa, LSTMs, TransformerXL, GPT-2 XL; see Section 3.3.1) are given in Figure 3.2. Correlations between model IC verb biases and human IC verb biases were calculated using Pearson's r.[13] As in the above section, there was variation in the degree that models match

[13]We used the implementation of Pearson's r from scipy.

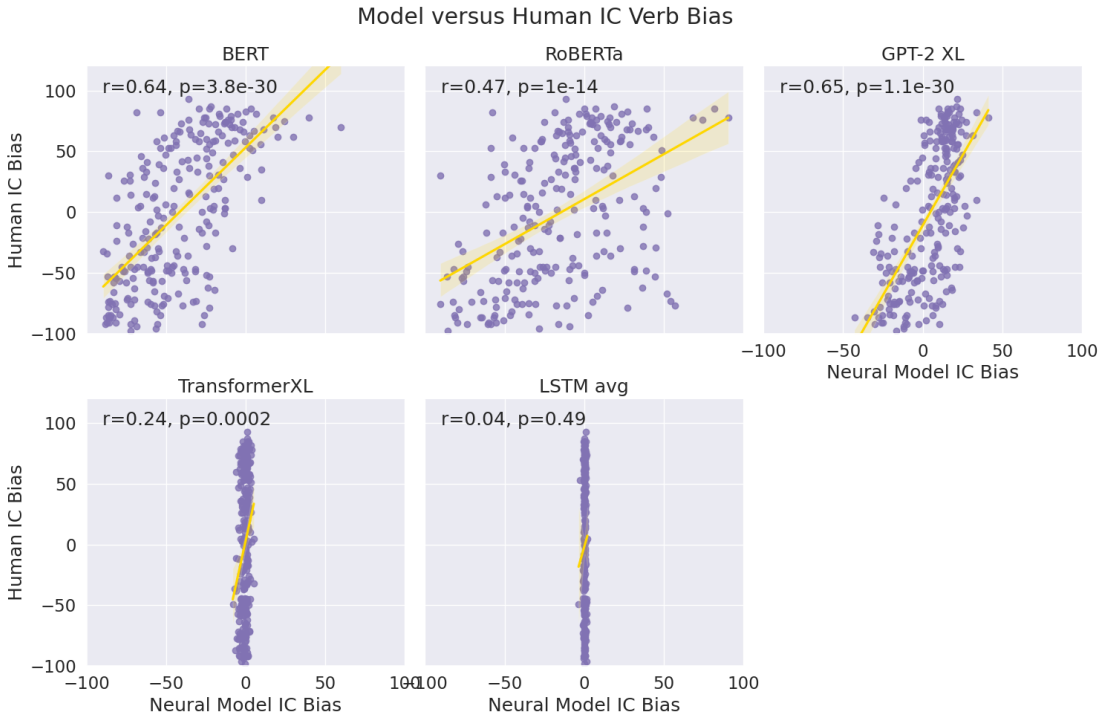

Figure 3.2: Correlation between human and model IC verb bias. Human biases are from Ferstl et al. (2011). Model biases are the scaled average difference between the probability of pronouns referring to the subject and pronouns referring to the object (see Section 3.3.3 for more details). A value of 100 corresponds to a verb with a complete subject-bias, and a value of -100 to a verb with a complete object-bias.

human IC biases. LSTMs showed no significant correlation with human IC bias, while TransformerXL showed a moderate correlation with human IC biases. The remaining transformers showed strong correlations with human IC bias, with GPT-2 XL and BERT achieving the best fit. As can be seen with Table 3.1, even models with strong correlations with the results in Ferstl et al. (2011) differ from humans in their specific IC biases (e.g., some object-biased IC verbs behaved like subject-biased IC verbs for neural models). Instead, neural models capture broader trends of gradient IC biases observed in humans.

	Bias	Human	BERT	RoBERTa	GPT-2 XL
1	Object	valued	venerated	treasured	despised
2	Object	laughed at	admired	*dreamed about	praised
3	Object	congratulated	disliked	cherished	mocked
4	Object	admired	deplored	valued	commended
5	Object	noticed	liked	prized	admired
6	Object	thanked	esteemed	relished	derided
7	Object	carried	adored	feared	decried
8	Object	liked	feared	appreciated	disliked
9	Object	hated	detested	liked	rebuked
10	Object	respected	*dreamed about	wanted	lauded
1	Subject	apologized to	unnerved	captivated	lied to
2	Subject	attracted	captivated	fascinated	fooled
3	Subject	agitated	unsettled	intrigued	telephoned
4	Subject	delighted	infuriated	*comforted	deceived
5	Subject	fascinated	astounded	*guided	*filmed
6	Subject	angered	enraged	*uplifted	confessed to
7	Subject	pleased	inspired	*counseled	cheated
8	Subject	called	amazed	astonished	confided in
9	Subject	telephoned	startled	affected	*played with
10	Subject	charmed	surprised	*supported	hurt

Table 3.1: Top 10 most Object and Subject-biased IC verbs for Humans (from Ferstl et al., 2011), BERT, RoBERTa, and GPT-2 XL. An asterisk denotes verbs which have the opposite qualitative bias for humans (e.g., *comforted* is object-biased for humans).

3.3.6 Discussion

Above we examined the extent to which discourse structure, determined by implicit causality verbs, could be acquired by transformer and LSTM neural models of language (cf. *Sally frightened Mary because she...* and *Sally feared Mary because she...*). Specifically, we evaluated, via comparison to human experiments, whether IC verb biases could influence the probability of pronouns. Given the claims in recent literature that implicit causality arises without extra pragmatic inference on the part of human comprehenders, we hypothesized that neural models would be able to acquire such contrasts (analogous to their ability to acquire syntactic

agreement).

We found that LSTM models were unable to demonstrate knowledge of IC in predicting pronouns. However, a transformer (TransformerXL) trained on the exact same data as the LSTM models was able to partially represent an IC distinction. In evaluating transformer models trained on vastly more data (GPT-2 XL, BERT, RoBERTa), we found a more robust, human-like sensitivity to IC bias when predicting pronouns: subject-biased IC verbs increased model preference for pronouns referring to the subject and object-biased IC verbs increased model preferences for pronouns referring to the object.

From a theoretical perspective, the findings provide additional support for the centering of implicit causality within the linguistic signal proper. That is, IC bias is learnable, to some degree, without pragmatic inference as hypothesized in Section 3.3 (see also Hartshorne, 2014). For the transformer models that did learn aspects of the human-like IC verb biases, they captured both the categorical distinction between subject and object-biased verbs, and also the gradient bias humans associate with the verbs.

In fact, the transformer IC verb bias does in fact seem to be grounded in the semantics of the verb, rather than just picking out the surface subject and the surface object. Consider the interaction of IC verb bias and passivization:

(4) a. The man admired the woman because ...

b. The woman was admired by the man because ...

For (4), English speakers prefer continuations that refer to the woman (i.e. predicting the pronoun "she"), despite *woman* being the surface object in (4-a) and the surface

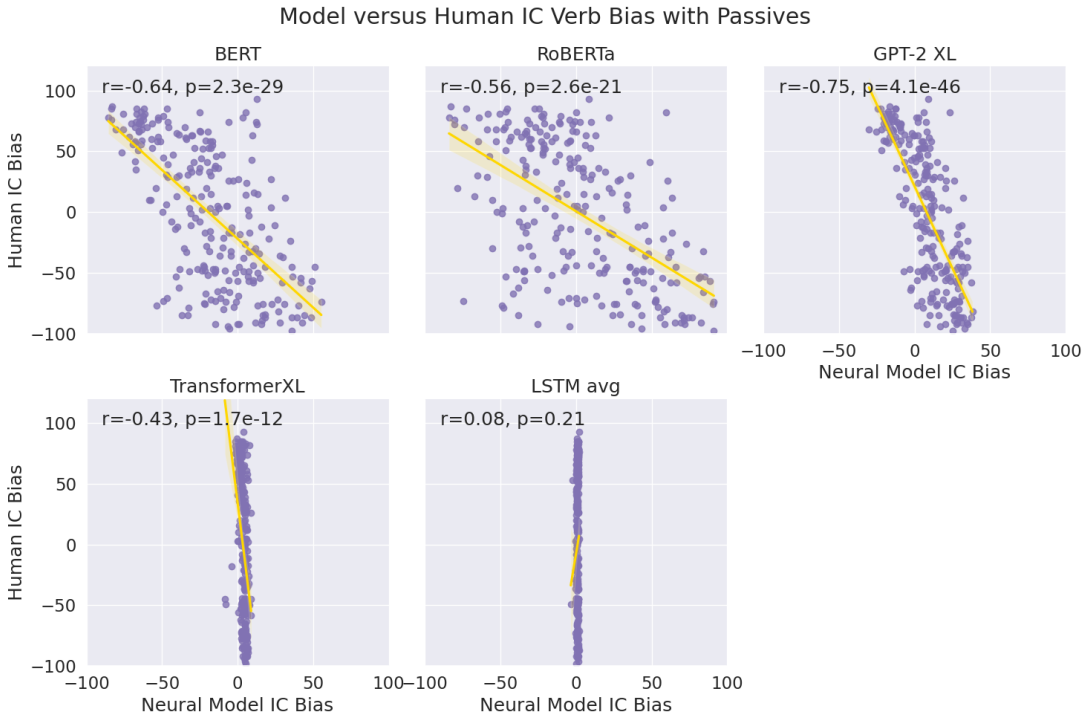

Figure 3.3: Correlation between human and model IC verb bias. Human biases are from Ferstl et al. (2011). Model biases are the scaled average difference with the passive construction between the probability of pronouns referring to the subject and pronouns referring to the object (see Section 3.3.3 for more details). A value of 100 corresponds to a verb with a complete subject-bias, and a value of -100 to a verb with a complete object-bias.

subject in (4-b) (for discussion see Garvey et al., 1974). That is, IC verb bias selects for a thematic role (e.g., the experiencer), rather than strictly a surface position (see Hartshorne et al., 2013). Transformer models also captured this distinction, flipping the predicted pronoun when the verb was passivized. In fact, the correlation between human IC bias (with active voice) and the model IC bias (with passive voice) was stronger (GPT-2 XL had a correlation of 0.64 with active constructions and 0.75 for passive constructions; see Figure 3.3).

Thus, it appears that transformer models learn a relatively abstract implicit

causality verb bias that is in line with results from human psycholinguistic studies. In the following sections, we explored the interactions between implicit causality and other linguistic processes cross-linguistically.

3.4 Cross-linguistic Instability of IC in Neural Models of Language

In the previous sections, we found evidence that transformer neural models of language learn IC biases in predicting pronouns. Thus, we might be cautiously optimistic about the ability of neural models to acquire an IC bias from just text. The ability of models trained on other languages to acquire an IC bias, however, has not been explored, and we turn now to cross-linguistic comparison.

Within the psycholinguistic literature, IC has been shown to be remarkably consistent cross-linguistically (see Hartshorne et al., 2013; Ngo and Kaiser, 2020). That is, IC verbs have been attested in a variety of languages. Given the cross-linguistic consistency of IC, then, models trained on other languages should also demonstrate an IC bias. However, using two popular model types, BERT based (Devlin et al., 2019) and RoBERTa based (Liu et al., 2019),[14] we find that models only acquired a human-like IC bias in English and Chinese, but not in Spanish and Italian.

We relate this to a crucial difference in the presence of a competing linguistic constraint affecting pronouns in the target languages. Spanish and Italian have

[14]These model types were chosen for ease of access to existing models. Pre-trained, large auto-regressive models are largely restricted to English, moreover Section 3.3 suggested that LSTMs are limited in their ability to acquire an IC bias in English (see also Davis and van Schijndel, 2020a).

a well studied process *pro drop*, which allows for subjects to be 'empty' (Rizzi, 1986). An English equivalent would be "(she) likes BERT" where *she* can be elided. While IC verbs increase the probability of a pronoun that refers to a particular antecedent, pro drop disprefers any overt pronoun in subject position (i.e. the target location in our study). That is, both processes are in direct competition in our experiments. As a result, Spanish and Italian models are susceptible to overgeneralizing any learned pro drop knowledge, favoring no pronouns rather than IC-conditioned pronoun generation.

Therefore to exhibit an IC bias, models of Spanish and Italian have two tasks: learn the relevant constraints (i.e. IC and pro drop) and the relative ranking of these constraints. We find that neural models learn both constraints, but, critically, instantiate the wrong ranking, favoring pro drop to an IC bias. Using fine-tuning to demote pro drop, we are able to uncover otherwise dormant IC knowledge in Spanish and Italian. Thus, the apparent failure of the Spanish and Italian models to pattern like English and Chinese is not evidence on its own of a model's inability to acquire the requisite linguistic knowledge, but is in fact evidence that models are unable to adjudicate between competing linguistic constraints in a human-like way. In English and Chinese, the promotion of a pro drop process via fine-tuning has the opposing effect, diminishing an IC bias in model behavior. As such, our results suggest mismatches in either the learning of linguistic constraints or their relative ranking induce non-human like behavior.

3.4.1 Methods: Neural Models of Language

Due to the lack of pre-trained autoregressive models beyond English, we focused on two popular non-autoregressive language model variants, BERT (Devlin et al.,

Model	Lang	Tokens
BERT	EN	3.3B
RoBERTa	EN	30B
Chinese BERT	ZH	5.4B
Chinese RoBERTa	ZH	5.4B
BETO	ES	3B
RuPERTa	ES	3B
Italian BERT	IT	2B
UmBERTo	IT	0.6B
GilBERTo	IT	11B

Table 3.2: Summary of models investigated with language and approximate number of tokens in training. For RoBERTa we use the approximation given in Warstadt et al. (2020b).

2019) and RoBERTa (Liu et al., 2019), which have variants trained on several different languages. We used existing models available via HuggingFace (Wolf et al., 2020).

We found that mBERT exhibited no IC bias in English, confirming existing work that has found multilingual models perform worse than monolingual models on targeted linguistic tasks (e.g., Mueller et al., 2020).[15] Therefore, we only investigated monolingual models (summarized in Table 3.2). For English, we used the BERT base uncased model and the RoBERTa base model explored in Section 3.3. For Chinese, we evaluated BERT and RoBERTa models from Cui et al. (2020). For Spanish, we used BETO (Cañete et al., 2020) and RuPERTa (Romero, 2020). For Italian, we evaluated an uncased Italian BERT[16] as well as two RoBERTa based models, UmBERTo (Parisi et al., 2020) and GilBERTo (Ravasio and Di Perna, 2020).

[15]Results are provided in Appendix A.2.

[16]https://huggingface.co/dbmdz/bert-base-italian-uncased

3.4.2 Methods: Stimuli

Our list of target verbs was derived from existing psycholinguistic studies of IC verbs.[17] For English, we used the IC verbs from Ferstl et al. (2011). Each verb in the human experiment was coded for IC bias based on continuations of sentence fragments (e.g., *Kate accused Bill because ...*). For Spanish, we used the IC verbs from Goikoetxea et al. (2008), which followed a similar paradigm as Ferstl et al. (2011) for English. Participants were given sentence fragments and asked to complete the sentence and circle their intended referent. The study reported the percent of subject continuations for 100 verbs, from which we used the 61 verbs which had a significant IC bias (i.e. we excluded verbs with no significant subject or object bias).

For Italian, we used the 40 IC verbs reported in Mannetti and De Grada (1991). Human participants were given ambiguous completed sentences with no overt pronoun like "John feared Michael because of the kind of person (he) is" and asked to judge who the null pronoun referred to, with the average number of responses that gave the subject as the antecedent reported.[18] For Chinese, we used 59 IC verbs reported in Hartshorne et al. (2013), which determined average subject bias per verb in a similar way as Mannetti and De Grada (1991) (i.e. judgments of antecedent preferences given ambiguous sentences containing with overt pronouns).[19]

[17]For each language investigated, the stimuli were evaluated for grammaticality by native speakers with academic training in linguistics. All IC verbs and noun pairs are given in Appendix A.1.

[18]Specifically, Mannetti and De Grada (1991) grouped the verbs into four categories and reported the average per category as well as individual verb results for the most biased verbs and the negative/positive valency verbs. Additionally, figures showing average responses across various conditions was reported for one of the categories. From the combination of this information, the average scores for all but two verbs were determined. The remaining two verbs were assigned the reported average score of its stimuli group.

[19]In Hartshorne et al. (2013), 60 verbs were reported, but after consultation with a native speaker

We generated stimuli using 14 pairs of stereotypical male and female nouns (e.g., *man* vs. *woman*, *husband* vs. *wife*) in each language, rather than rely on proper names as was done in the human experiments. The models we investigated are bidirectional, so we used a neutral right context, *was there*, for English and Spanish, where human experiments provided no right context.[20] For Italian we utilized the full sentences investigated in the human experiments. The Chinese human experiment also used full sentences, but relied on nonce words (i.e. novel, constructed words like *sliktopoz*), so we chose instead to generate sentences like the English and Spanish ones. All stimuli had subjects and objects that differed in gender, such that all nouns occurred in subject or object position (i.e. the stimuli were fully balanced for gender):

(5) the man admired the woman because [MASK] was there.[21]

Additionally, gender agreement in Spanish and Italian allowed us to investigate the ability of models to apply knowledge of implicit causality when no overt pronoun was given (a crucial distinction we return to again in Section 3.4.5). The stimuli were similar to the above set, however no pronoun was included and the final position (meant for an adjective) was masked:

(6) el hombre despreció a la mujer porque estaba [MASK].

In (6), the gender on the adjectives at the mask location pick out either the subject

with academic training in linguistics, one verb was excluded due to perceived ungrammaticality of the construction.

[20]Using *here, outside,* or *inside* as the right context produces qualitatively the same patterns.

[21]The model specific mask token was used. Additionally, all models were uncased, with the exception of RoBERTa, so lower cased stimuli were used.

(which is masculine) or the object (which is feminine).

3.4.3 Methods: Measures

As in Section 3.3.3, the mismatch in gender between the subject and the object forced the choice of pronoun to be unambiguous. For the pronoun stimuli, we gathered the probability assigned to the third person singular male and female pronouns (e.g., *he* and *she*).[22] Our measures were grouped by antecedent type (i.e. the pronoun refers to the subject or the object) and whether the verb was object-biased or subject-biased. For example, BERT assigns to (5) a score of 0.01 for the subject antecedent (i.e. *he*) and 0.97 for the object (i.e. *she*), in line with the object-bias of *admire*.

For the Spanish and Italian adjectival stimuli (e.g., (6)), we gathered the 100 most likely continuations and parsed them using spacy. Gendered adjectives unambiguously referred to either the subject or the object, as in the case of pronouns, so we grouped our results by antecedent type and the IC bias of the verb. For example, BETO (Spanish BERT) assigned to (6) a probability of 0.65 to adjectives with feminine gender and a probability of 0.02 to adjectives with masculine gender, in line with the object-bias of *despreció* (despised).

3.4.4 Models Inconsistently Capture Implicit Causality

As exemplified in (1), repeated below as (7), IC verb bias modulates the preference for pronouns.

[22]In spoken Chinese, the male and female pronouns are homophones. They are, however, distinguished in writing.

(7) a. Lavender frightened Kate because she was so terrifying.

b. Lavender admired Kate because she was so amazing.

An object-biased IC verb (e.g., *admired*) should increase the likelihood of pronouns that refer to the object, and a subject-biased IC verb (e.g., *frightened*) should increase the likelihood of reference to the subject. Given that all the investigated stimuli were disambiguated by gender, we categorized our results by the antecedent of the pronoun and the IC verb bias. We first turn to English and Chinese, which showed an IC bias in line with existing work on IC bias in autoregressive English models (e.g., Upadhye et al., 2020; Davis and van Schijndel, 2020a) We then detail the results for Spanish and Italian, where only very limited, if any, IC bias was observed.

English and Chinese

The results for English and Chinese are given in Figure 3.4 and detailed in Appendix A.2. All models demonstrated a greater preference for pronouns referring to the object after an object-biased IC verb than after a subject-biased IC verb.[23] Additionally, they had greater preferences for pronouns referring to the subject after a subject-biased IC verb than after a object-biased IC verb. That is, all models showed the expected IC-bias effect. Generally, there was an overall greater preference for referring to the object, in line with a recency bias, with the exception of RoBERTa, where subject-biased IC verbs neutralized the recency effect.

[23]Throughout the following sections, statistical significance was determined by two-way t-tests evaluating the difference between pronouns referring to objects after subject-biased and object-biased IC verbs, and similarly for pronouns referring to the subject. The threshold for statistical significance was $p = 0.0006$, after adjusting for the 88 statistical tests conducted in this section.

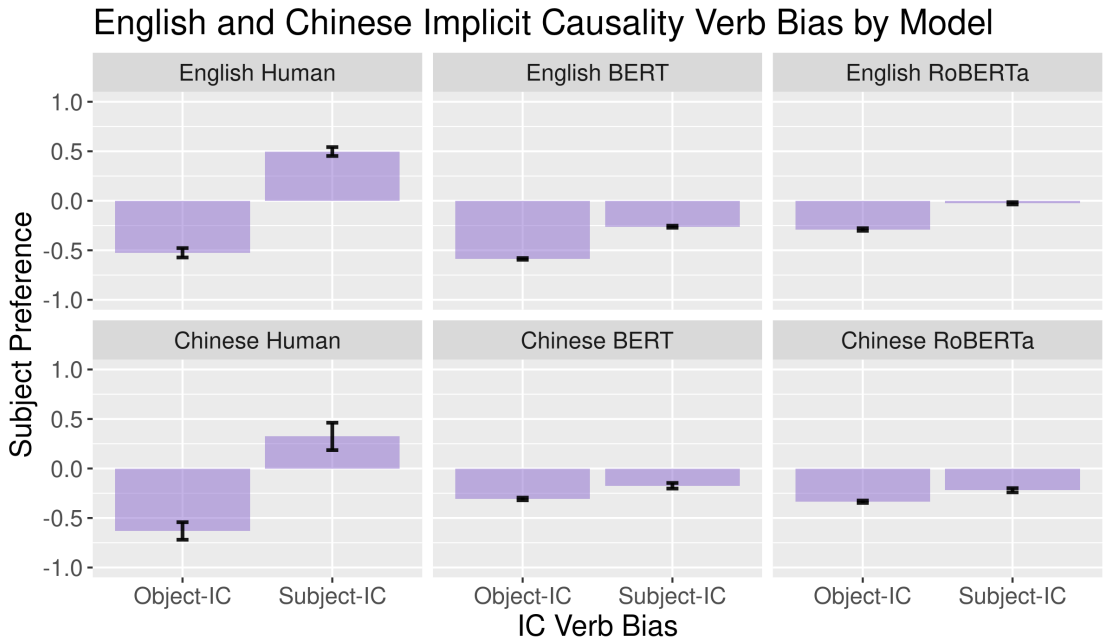

Figure 3.4: Subject preference grouped by implicit causality verb type for humans (English is from Ferstl et al., 2011; Chinese is from Hartshorne et al., 2013), English BERT and RoBERTa, and Chinese BERT and RoBERTa. A value of 1.0 corresponds to a complete preference for pronouns agreeing in gender with the subject (i.e. a subject bias), and a value of -1.0 corresponds to a complete preference for pronouns agreeing with the object (i.e. an object bias). Error bars are 95% confidence intervals.

Spanish and Italian

The results for Spanish and Italian are given in Figure 3.5 and detailed in Appendix A.2. In stark contrast to the models of English and Chinese, an IC bias was either not demonstrated or was only weakly attested. For Spanish, BETO showed a greater preference for pronouns referencing the object after an object-biased IC verb than after a subject-biased IC verb. There was no corresponding IC effect for pronouns referring to the subject, and RuPERTa (a RoBERTa based model) had no IC effect at all. Similarly, for the stimuli without pronouns, where gender agreement on the predicted adjective disambiguated the antecedent, there were no

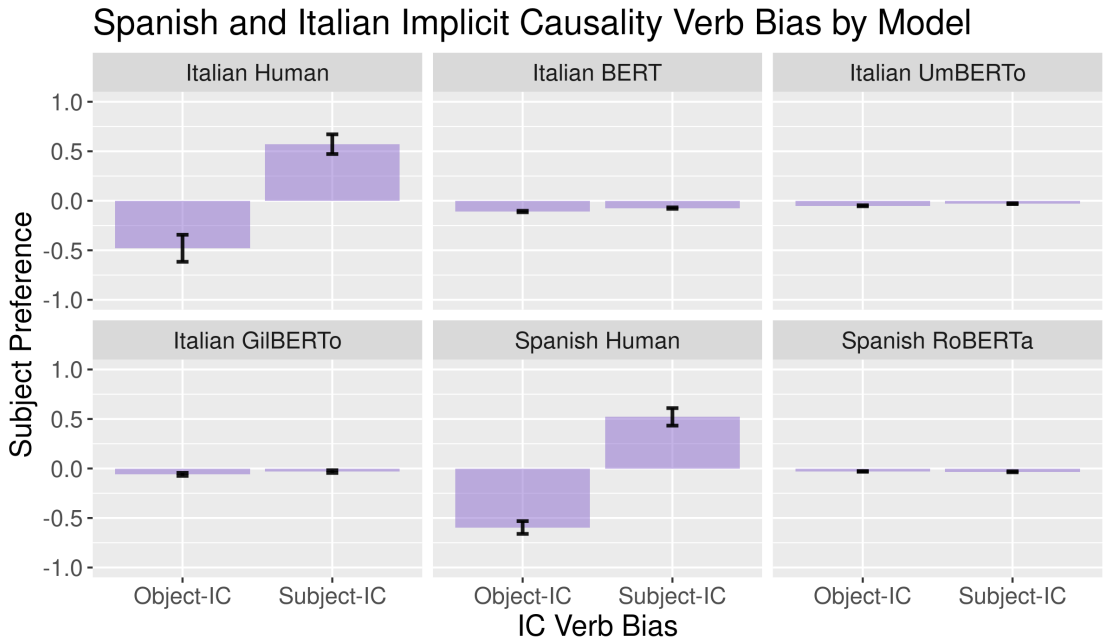

Figure 3.5: Subject preference grouped by implicit causality verb type for humans (Italian is from Mannetti and De Grada, 1991; Spanish is from Goikoetxea et al., 2008), Italian BERT, UmBERTo, and GilBERTo, and Spanish RoBERTa. A value of 1.0 corresponds to a complete preference for pronouns agreeing in gender with the subject (i.e. a subject bias), and a value of -1.0 corresponds to a complete preference for pronouns agreeing with the object (i.e. an object bias). Error bars are 95% confidence intervals.

significant interactions between IC and antecedent for Spanish.

Italian BERT and GilBERTo (a RoBERTa based model) had no significant effect of IC-verb on pronouns referring to the object. There was a significant, albeit very small increased probability for pronouns referring to the subject after a subject-biased IC verb in line with a weak subject-IC bias. Similarly, UmBERTo (a RoBERTa based model) had significant, yet tiny IC effects, where object-biased IC verbs increased the probability of pronouns referring to objects compared to subject-biased IC verbs (similarly with pronouns referring to the subject being more likely after subject-biased IC verbs as compared to object-biased IC verbs).

As with Spanish, for the stimuli without pronouns, where gender agreement on the predicted adjective disambiguated the antecedent, there were no significant interactions between IC and antecedent for Italian, though Italian BERT and UmBERTo had marginal effects that were in the expected directions: subject-biased IC verbs increased the likelihood of adjectives that agreed with subjects compared to object-biased IC verbs, and object-biased IC verbs increased the likelihood of adjectives that agreed with the object as compared to subject-biased IC verbs.

Any significant effects in Spanish and Italian were much smaller than their counterparts in English (as is visually apparent between Figure 3.4 and Figure 3.5), and each of the Spanish and Italian models failed to demonstrate at least one of the IC effects.

3.4.5 Competing Constraints: Pro Drop and Implicit Causality

We were left with an apparent mismatch between models of English and Chinese and models of Spanish and Italian. In the former, an IC verb bias modulated pronoun preferences. In the latter, the same IC verb bias was comparably absent. Recall that, for humans, the psycholinguistic literature suggests that IC bias is, in fact, quite consistent across languages (see Hartshorne et al., 2013).

Careful consideration of the languages under investigation suggests a possible reason for why the two sets of models behave so differently. Languages can be thought of as systems of competing linguistic constraints (e.g., Optimality Theory; Prince and Smolensky, 2004). Spanish and Italian exhibit pro drop. Typical

grammatical sentences often lack overt pronouns in subject position, opting instead to rely on rich agreement systems to disambiguate the intended subject at the verb (Rizzi, 1986). This constraint competes with IC, which favors pronouns that refer to either the subject or the object. Chinese also allows for empty arguments (both subjects and objects), typically called *discourse pro drop* (Huang, 1984).[24] As the name suggests, however, this process is more discourse constrained than the process in Spanish and Italian. For example, the empty subject can only refer to the subject of the preceding sentence (see Liu, 2014). As a means of comparison, in surveying three Universal Dependencies datasets,[25] 8% of nsubj (or nsubj:pass) relations were pronouns for Chinese, while only 2% and 3% were pronouns in Spanish and Italian respectively. Finally, English lies on the opposite end of the continuum, requiring overt pronouns in the absence of other nominals (cf. *He likes NLP* and **Likes NLP*).

Therefore, it is possible that the presence of competing constraints in Spanish and Italian obscured any underlying IC knowledge: one constraint preferring pronouns which referred to the subject or object and the other constraint penalizing overt pronouns in subject positions (i.e. the target position masked in our experiments). In the following sections, we removed, or otherwise demoted, the dominance of each model's pro drop constraint for Spanish and Italian, and introduced, or promoted, a pro drop like constraint in English and Chinese. We found that the degree of IC bias in model behavior could be controlled by the presence, or absence, of a competing pro drop constraint.

[24]Other names common in the literature include *topic drop*, *radical pro drop*, and *rampant pro drop*.

[25]Chinese GSD, Italian ISDT, and Spanish AnCora.

Methodology

We constructed two classes of data to fine-tune the models on. The first aimed to demote the pro drop constraint in Spanish and Italian. The second aimed to inject a pro drop constraint into English and Chinese. For both we relied on Universal Dependencies datasets. For Spanish, we used the AnCora Spanish newswire corpus (Taulé et al., 2008), for Italian we used ISDT[26] and VIT (Delmonte et al., 2007), for English we used the English Web Treebank (Silveira et al., 2014), and for Chinese, we used the Traditional Chinese Universal Dependencies Treebank[27] annotated by Google (GSD) and the Chinese Parallel Universal Dependencies (PUD) corpus from the 2017 CoNLL shared task.[28]

For demoting pro drop, we found finite (i.e. inflected) verbs that did not have a subject relation in the corpora.[29] We then added a pronoun, matching the person and number information given on the verb, alternating the gender. For Italian, this amounted to a dataset of 3798 sentences with a total of 4608 pronouns (2,284 he or she) added. For parity with Italian, we restricted Spanish to a dataset of the first 4000 sentences, which had 5,559 pronouns (3,573 he or she) added. For the addition of a pro drop constraint in English and Chinese, we found and removed pronouns that bore a subject relation to a verb. This amounted to 935 modified sentences and 1083 removed pronouns (774 he or she) in Chinese and 6871 modified sentences and 10386 removed pronouns (2475 he or she) in English.[30]

For each language, 500 unmodified sentences were used for validation, and

[26] `https://github.com/UniversalDependencies/UD_Italian-ISDT`
[27] `https://github.com/UniversalDependencies/UD_Chinese-GSD`
[28] `https://github.com/UniversalDependencies/UD_Chinese-PUD`
[29] In particular, verbs that lacked any nsubj, nsubj:pass, expl, expl:impers, or expl:pass dependents
[30] A fuller breakdown of the fine tuning data is given in Appendix A.3.

unchanged versions of all the sentences were kept and used to fine-tune the models as a baseline to ensure that there was nothing about the data themselves that changed the IC-bias of the models. Moreover, we filtered the data to ensure that no verbs evaluated in our test data were included. Fine-tuning proceeded using HuggingFace's API. Each model was fine-tuned with a masked language modeling objective for 3 epochs with a learning rate of 5e-5, following the fine-tuning details in (Devlin et al., 2019).

Demoting Pro Drop: Spanish and Italian

As a baseline, we fine-tuned the Spanish and Italian models on unmodified versions of all the data we used for demoting pro drop. We found the same qualitative effects detailed in above, confirming that the data used for fine tuning when unmodified did not result in IC biased model behavior.

We turn now to our main experimental manipulation, fine-tuning the Spanish and Italian models on sentences that do not exhibit a pro drop process. It is worth repeating that the fine-tuning data shared no verbs or sentence frames with our test data. The results are given in Figure 3.6. Strikingly, an object-biased IC effect (pronouns referring to the object were more likely after object-biased IC verbs than subject-biased IC verbs) was observed for Italian BERT and GilBERTo despite no such effect being observed in the base models. Moreover, both models showed a more than doubled subject-biased IC verb effect. UmBERTo also showed increased IC effects, as compared to the base models. Similarly for Spanish, a subject-biased IC verb effect materialized for BETO when no corresponding effect was observed with the base model. The object-biased IC verb effect remained similar to what was reported in above. For RuPERTa, which showed no IC knowledge in the initial

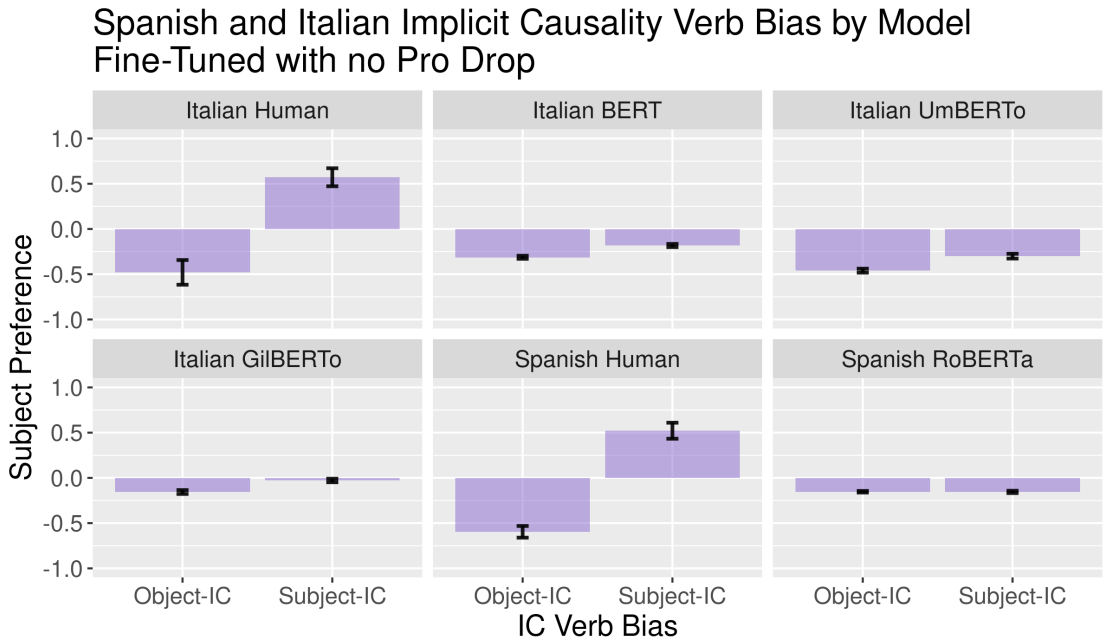

Figure 3.6: Subject preference after fine-tuning on sentences removing pro drop (i.e. adding a subject pronoun). Results are grouped by implicit causality verb type for humans (Italian is from Mannetti and De Grada, 1991; Spanish is from Goikoetxea et al., 2008), Italian BERT, UmBERTo, and GilBERTo, and Spanish RoBERTa. A value of 1.0 corresponds to a complete preference for pronouns agreeing in gender with the subject (i.e. a subject bias), and a value of -1.0 corresponds to a complete preference for pronouns agreeing with the object (i.e. an object bias). Error bars are 95% confidence intervals.

investigation, no IC knowledge surfaced after fine-tuning. We take this to suggest that RuPERTa has no underlying knowledge of IC, though further work should investigate this claim.

Taken together these results suggest that simply fine-tuning on a small number of sentences can re-rank linguistic constraints influencing model behavior and uncover other linguistic knowledge (in our case an underlying IC-bias). That is, model behavior was not necessarily incorrect or succumbing to some non-linguistic

bias in our initial exploration, but the models had in fact over-zealously learned one narrow aspect of the linguistic structure at the expense of another.

Further evidence that the re-ranking is targeting the interaction of pro-drop and IC-bias was found in evaluating the fine-tuned Spanish and Italian models on the stimuli which had no overt pronoun (i.e. the stimuli where the predicted agreement on the adjective disambiguated the antecedent). There was no change in the IC bias after fine-tuning for these conditions. That is the Spanish and Italian neural models continued to show no evidence of IC bias influencing gender agreement on the adjective.

Promoting Pro Drop: English and Chinese

In seeking to solidify the role of pro drop in obscuring underlying knowledge of IC, we turn to fine-tuning a pro drop constraint into models of English and Chinese. Recall that both models showed an IC effect, for both object-biased and subject-biased IC verbs. Moreover, both languages lack the pro drop process found in Spanish and Italian (though Chinese allows null arguments).

As with Spanish and Italian, we fine-tuned the English and Chinese models on unmodified versions of the training sentences as a baseline (i.e. the sentences kept their pronouns). There was no qualitative difference from the IC effects noted above. That is, for both English and Chinese, pronouns referring to the object were more likely after object-biased IC verbs than after subject-biased IC verbs and conversely pronouns referring to the subject were more likely after subject-biased than object-biased IC verbs.

The results after fine-tuning the models on data mimicking a Spanish and Italian

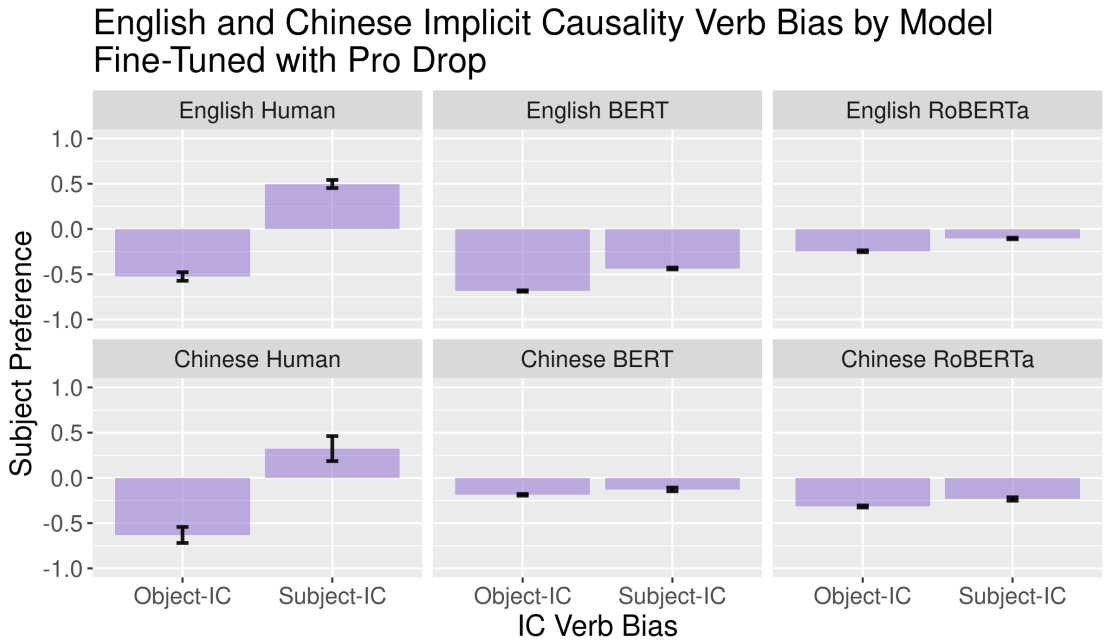

Figure 3.7: Subject preference after fine-tuning on sentences with pro drop (i.e. no subject pronoun). Results are grouped by implicit causality verb type for humans (English is from Ferstl et al., 2011; Chinese is from Hartshorne et al., 2013), English BERT and RoBERTa, and Chinese BERT and RoBERTa. A value of 1.0 corresponds to a complete preference for pronouns agreeing in gender with the subject (i.e. a subject bias), and a value of -1.0 corresponds to a complete preference for pronouns agreeing with the object (i.e. an object bias). Error bars are 95% confidence intervals.

like pro drop process (i.e. no pronouns in subject position) are given in Figure 3.7 and detailed in Appendix A.2. Despite fine-tuning on only 0.0004% and 0.003% of the data RoBERTa and BERT were trained on, respectively, the IC effects observed in above were severely diminished in English. However, the subject-biased IC verb effect remained robust in both models. For Chinese BERT, the subject-biased IC verb effect in the base model was lost and the object-biased IC verb effect was reduced. The subject-biased IC verb effect was similarly attenuated in Chinese RoBERTa. However, the object-biased IC verb effect remained.

For both languages, the IC effect was weakened, and even removed in the case of

subject-biased IC verbs in Chinese BERT, with relatively little evidence of pro drop. This result strengthens our claim that competition between linguistic constraints can obscure underlying linguistic knowledge in model behavior.

3.4.6 Discussion

This section investigated the ability of RoBERTa and BERT models to demonstrate knowledge of implicit causality across four languages (recall the contrast between *Lavender frightened Kate* and *Lavender admired Kate* in (1)). Contrary to humans, who show consistent subject and object-biased IC verb preferences across languages (see Hartshorne et al., 2013), BERT and RoBERTa models of Spanish and Italian failed to demonstrate the full IC bias found in English and Chinese BERT and RoBERTa models (with our English results supporting prior work on IC bias in neural models and extending it to non-autoregressive models; Upadhye et al., 2020; Davis and van Schijndel, 2020a). Following standard behavioral probing (e.g., Linzen et al., 2016), this mismatch may have been taken as evidence of differences in linguistic knowledge across languages. That is, model behavior in Spanish and Italian was inconsistent with predictions from the psycholinguistic IC literature, suggesting that these models lack knowledge of implicit causality. However, we found that to be an incorrect inference; the models *did* have underlying knowledge of IC.

Other linguistic processes influence pronouns in Spanish and Italian, and we showed that competition between multiple distinct constraints affects model behavior. One constraint (pro drop) decreases the probability of overt pronouns in subject position, while the other (IC) increases the probability of pronouns that refer to particular antecedents (subject-biased verbs like *frightened* favoring

subjects and object-biased verbs like *admired* favoring objects). Models of Spanish and Italian, then, must learn not only these two constraints, but also their ranking (i.e. should the model generate a pronoun as IC dictates, or generate no pronoun in line with pro drop). By fine-tuning the models on data contrary to pro drop (i.e. with overt pronouns in subject position), we uncovered otherwise hidden IC knowledge. Moreover, we found that fine-tuning a Spanish and Italian-like pro drop constraint into English and Chinese could greatly diminish IC's influence on model behavior (with as little as 0.0004% of a models original training data).

Taken together, we conclude that there are two ways of understanding mismatches between model linguistic behavior and human linguistic behavior. Either a model fails to learn the necessary linguistic constraint, or it succeeds in learning the constraint but fails to learn the correct interaction with other constraints. Existing literature points to a number of reasons a model may be unable to learn a linguistic representation, including the inability to learn mappings between form and meaning and the lack of embodiment (e.g., Bender and Koller, 2020; Bisk et al., 2020). We suggest that researchers should re-conceptualize linguistic inference on the part of neural models as inference of constraints and constraint ranking rather than as inference of symbolic linguistic knowledge in order to better understand model behavior. We believe such framing will open additional connections with linguistic theory and psycholinguistics. Minimally, we believe targeted fine-tuning for constraint re-ranking may provide a general method both to understand what linguistic knowledge these models possess and to aid in making their linguistic behavior more human-like.

3.5 General Discussion

Neural models of language have been claimed to learn, at least some, aspects of syntactic knowledge. The successes of transformer models to capture human-like IC biases in pronoun prediction extend these successes to discourse structure. However, neural models are unable to learn robust, human-like linguistic systems involving IC biases. That is, models fail to learn the interaction between IC and ambiguous relative clause attachment in English. Moreover, the successes in predicting pronouns in line with IC verb biases does not extend to all languages. The presence of competing processes targeting pronouns in certain languages (like Spanish and Italian) obscured underlying knowledge of IC. Models, then, appear able to learn individual constraints, but struggle in ranking constraints in a human-like fashion.

While the origin of this mismatch is left to ongoing work, I suspect the mismatch follows from the learning objective of such models, coupled with the prevalence of the relevant processes in the language. These neural models are trained to optimize the objective of predicting the next word (or some word in context for non-autoregressive models). Thus, if two processes have conflicting influences for the same word, it would seem that the process that occurs more often in the data wins out.

For concreteness, suppose the neural model of language is trying to learn to weight the constraints NoPronoun and ICBias. NoPronoun targets pronouns and assigns negative weight to continuations that have an overt pronoun. ICBias targets pronouns and assigns positive weight to continuations that have an overt pronoun agreeing the IC bias of the verb. The contexts which trigger ICBias are a subset of those that trigger NoPronoun, as contexts with IC verbs are a subset of all possible

contexts which support pronouns. Suppose that learning proceeds by weighting the constraints in accordance to their frequency of occurrence, so that every context without a pronoun increments the NoPronoun constraint, and every context with a pronoun agreeing with the IC verb bias increments the ICBias constraint (as with the Gradual Learning Algorithm for learning OT Grammers; see Boersma and Hayes, 2001). If the number of contexts where NoPronoun occurs is greater than the number of contexts where ICBias occurs (with an overt pronoun), then the NoPronoun constraint will outpace the ICBias constraint leading to a model which will avoid generating pronouns across the board.

This would additionally explain the lack of an interaction between IC and ambiguous relative clause attachment in English documented in the following chapter, even in cases where models did learn relative clause attachment (cf. Chapter 4), because the general low attachment bias would dominate the more specific high attachment bias of object-biased IC verbs. In fact, mismatches of this type are a general problem for gradual learning with weighted OT grammars suggesting that if neural models learn in this fashion, then they will generally fail in contexts where a more general constraint targets the same output that a more specific constraint governs (see Tessier, 2009).

Ultimately, this chapter has shown that investigations of isolated linguistic processes should be coupled with investigations of their interaction with other processes. In doing so, it appears that linguistic knowledge is overestimated by the minimal pair approach advocated by targeted syntactic evaluations. By centering interaction, we can find evidence for a broad class of phenomena that models will fail to capture. Namely, the interaction of general and specific constraints.

CHAPTER 4

AMBIGUOUS RELATIVE CLAUSE ATTACHMENT

4.1 Introduction

A growing body of research has evaluated whether large pre-trained language models contain linguistic knowledge (e.g., subject-verb agreement).[1] Often the evaluation of neural models proceeds by focusing on comparisons to human linguistic behaviors (e.g., acceptability judgments, reading times, comprehension questions). One dominate approach constructs minimal pairs that differ in some critical way, following existing studies of humans (e.g., the same sentence with grammatical or ungrammatical agreement).

While focusing on the behavior of humans and models may allow for some degree of separation from certain theoretical commitments (e.g., the exact mechanism underlying the behavior is only relevant in so far as the particular behavioral experiment exposes the mechanism), interpreting the results of these comparisons relies on certain assumptions (for further elaboration see Chapter 2). The following chapter challenges one of these basic assumptions: the necessary linguistic biases are in the training signal. That is, in order to interpret the success of models in mimicking human behaviors, we have to assume that a model could *reasonably* learn the behavior. This chapter casts doubt on the notion that the necessary linguistic biases for acquiring human interpretations preferences are present in the training signal at all. Thus, a model which fully, and only, mimics the biases that

[1]Code for replicating the experiments, figures, and statistical models in this chapter can be found on Github at `https://github.com/forrestdavis/Dissertation/tree/main/Attachment`. Templates for the stimuli are provided in Appendix B for ease of reference. Parts of this chapter appear in Davis and van Schijndel (2020c).

underlie the training data will fall short of human language comprehension. The failure of a model to match humans, in other words, may follow from a limitation of data, not a limitation of models.[2]

That does not preclude the possibility of a certain model, perhaps with inherent structure, from achieving human-like comprehension. Claiming that for a certain phenomenon it is impossible for a model to learn it is a strong stance. Moreover, it is a stance that seems trivially false because humans come to obtain comprehension biases, and certainly humans learn, at least something, from their primary linguistic experience (we do not all speak the same language). Instead, the present chapter's focus is narrowed to current models and training techniques, where we ask whether such neural models arrive at human comprehension biases, and if not, what might be driving this mismatch.

We utilize the, now common, evaluation technique of checking whether a model assigns higher probability to grammatical sentences compared to ungrammatical sentences (Linzen et al., 2016). However, we extend beyond binary grammaticality. Human linguistic knowledge extends beyond knowing the difference between valid and invalid sentences; neural models must also be able to correctly prioritize simultaneous valid interpretations in a human-like way (Lau et al., 2017). In this chapter, we investigate whether neural networks can in fact prioritize simultaneous syntactic forms in a human-like way. In particular, we probe the biases of neural networks for ambiguous relative clause (RC) attachments, such as the following:

(1) Andrew had dinner yesterday with the nephew of the teacher *that was divorced*. (from Fernández, 2003)

[2]Of course the model could also be limited (for example, see Section 3.3 where LSTMs underperform as compared to transformer models).

In (1), there are two nominals (*nephew* and *teacher*) that are available for modification by the RC (*that was divorced*). We refer to attachment of the RC to the syntactically higher nominal (i.e. the nephew is divorced) as HIGH and attachment to the lower nominal (i.e. the teacher is divorced) as LOW.

As both interpretations are equally semantically plausible when no supporting context is given, we might expect that humans choose between HIGH and LOW at chance. However, it has been widely established that English speakers tend to interpret the relative clause as modifying the lower nominal more often than the higher nominal (i.e. they have a LOW bias;[3] Carreiras and Clifton, 1993; Frazier and Clifton, 1996; Carreiras and Clifton, 1999; Fernández, 2003). LOW bias is actually typologically less common than HIGH bias (Brysbaert and Mitchell, 1996). A proto-typical example of a language with HIGH attachment bias is Spanish (see Carreiras and Clifton, 1993, 1999; Fernández, 2003).

A growing body of literature has shown that English linguistic structures conveniently overlap with non-linguistic biases in neural language models leading to performance advantages for models of English, without such models being able to learn comparable structures in non-English-like languages (e.g., Dyer et al., 2019). Moreover neural models can exhibit a recency bias (Ravfogel et al., 2019), suggesting that one of these attachment types (LOW), will be more easily learned. Therefore, the models might appear to perform in a human-like fashion on English, while failing on the cross-linguistically more common attachment preference (HIGH) found in Spanish.

However, prior work has shown, via a synthetic language experiment, that recurrent neural network language models are capable of learning either type of

[3]We use "bias" throughout this chapter to refer to "interpretation bias." We will return to the distinction between production bias and interpretation bias in Section 4.8.

attachment (Davis and van Schijndel, 2020c). In fact, they may even have a slight high attachment preference for these constructions. In this chapter, we expand this investigation to transformer models (BERT, RoBERTa, and GPT-2 based models to be exact), first by investigating the ability of these neural models to replicate the interpretation preferences documented in Cuetos and Mitchell (1988) and Fernández (2003) via number agreement. We then take a more nuanced looked at the influence of properties in the complex noun phrase on attachment preferences in neural models,. Next, we investigated the interaction between attachment and implicit causality in English. Finally, we examined gender agreement in Spanish relative clauses.

Across experiments and neural models, human-like interpretation preferences failed to be consistently learned. For Spanish, number agreement in the relative clause favored low attachment, in contrast to the ultimate preference for high attachment interpretations demonstrated for these same stimuli (e.g., Fernández, 2003). Even for English, careful comparison between neural models and humans demonstrated that neural models over-emphasize the low attachment preferences, continuing to favor low attachment in cases where humans favor high attachment. Investigations of gender agreement in Spanish point to one possible answer to this mismatch between humans and neural models; neural models more closely resemble human early reading behaviors rather than human interpretation preferences.

Taken together these results raise broader questions regarding the relationship between comprehension (i.e. typical language model use cases) and production (which generates the training data for language models) and point to a deeper inability of neural models of language to learn aspects of linguistic structure from text data alone.

4.2 Background

Much recent work has probed neural models of language for their ability to represent syntactic phenomena. In particular, subject-verb agreement has been explored extensively (e.g., Linzen et al., 2016; Bernardy and Lappin, 2017; Enguehard et al., 2017) with results at human level performance in some cases (Gulordava et al., 2018). However, additional studies have found that the models are unable to generalize sequential patterns to longer or shorter sequences that share the same abstract constructions (van Schijndel et al., 2019). This suggests that the learned syntactic representations are very brittle.

Despite this brittleness, neural models of language have been claimed to exhibit human-like behavior when processing garden path constructions (van Schijndel and Linzen, 2018a; Futrell and Levy, 2019; Frank and Hoeks, 2019), reflexive pronouns and negative polarity items (Warstadt et al., 2020a), and center embedding and syntactic islands (Wilcox et al., 2019a,b). There are some cases, like coordination islands, where model behavior is distinctly non-human (see Wilcox et al., 2019b), but in general this literature suggests that neural models encode some type of abstract syntactic representation (e.g., Prasad et al., 2019). Thus far though, the linguistic structures used to probe neural models of language have often been those with unambiguously ungrammatical counterparts. This extends into the domain of semantics, where benchmarks like GLUE and SuperGLUE evaluate neural models for correct vs. incorrect interpretations on tasks targeting language understanding (Wang et al., 2018, 2019).

Some recent work has relaxed this binary distinction of correct vs. incorrect or grammatical vs. ungrammatical. Lau et al. (2017) correlate acceptability scores generated from a neural model to average human acceptability ratings, suggesting

that human-like gradient syntactic knowledge can be captured by such models. Futrell and Levy (2019) also look at gradient acceptability in both RNN neural models and humans, by focusing on alternations of syntactic constituency order (e.g., heavy NP shift, dative alternation). Their results suggest that RNN neural models acquire soft constraints on word ordering, like humans. However, the alternations in Futrell and Levy, while varying in their degree of acceptability, maintain the same syntactic relations throughout the alternation (e.g., *gave a book to Tom* and *gave Tom a book* both preserve the fact that *Tom* is the indirect object). This work expands this line of research by probing how neural models of language behave when multiple valid interpretations, crucially with different syntactic relations, are available within a single sentence. We find that neural models do not consistently resolve such ambiguity in a human-like way.

4.3 Neural Models

We analyzed both long short-term memory networks (LSTMs; Hochreiter and Schmidhuber, 1997) and transformer models throughout the present chapter. For English, we used the 25 LSTM models trained on Wikitext-103 (Merity et al., 2016) that were detailed in Chapter 3. For transformers, we evaluated BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and GPT-2 XL Radford et al. (2019) via HuggingFace (Wolf et al., 2020).

For Spanish, five LSTMs were trained on Spanish Wikipedia data following the process used by Gulordava et al. (2018). A recent dump of Spanish Wikipedia was downloaded, text was extracted using WikiExtractor,[4] and tokenization was done

[4] `https://github.com/attardi/wikiextractor`

using TreeTagger. A 100-million word subset of the data was extracted, shuffled by sentence, and split into training (80%) and validation (10%) sets. For training, we included the 50K most frequent words in the vocabulary, replacing the other tokens with '⟨UNK⟩'.[5] For transformers, we used a BERT based model (BETO; Cañete et al., 2020), a RoBERTa based model (RuPERTa; Romero, 2020), and two GPT-2 based Spanish models, Spanish GPT-2 trained on the same corpus as BETO and GPT-2 Spanish trained on Wikipedia and books, all via HuggingFace (Wolf et al., 2020).[6]

4.4 Neural Models and Attachment Preferences

As detailed above, ambiguous attachment preferences differ across speakers depending on their languages. In Cuetos and Mitchell (1988), participants were given stimuli like *The journalist interviewed the daughter of the colonel who had the accident* and asked comprehension questions like *Who had the accident?*. Spanish participants favored answers following from high attachment (e.g., the daughter had the accident) approximately 60% of the time. English speakers, in contrast, favored answers following from low attachment with high attachment answers only occurring 37% of the time.[7] Additional experiments for Spanish found evidence for a high attachment preference in self-paced reading times.[8]

Additional studies have provided evidence for Spanish speakers high attachment

[5]More details are given in Davis and van Schijndel (2020c).

[6]The models were downloaded from Hugging Face's model hub at `https://huggingface.co/mrm8488/spanish-gpt2` and `https://huggingface.co/DeepESP/gpt2-spanish`, respectively.

[7]Cuetos and Mitchell (1988) focus their analysis on the subset of stimuli were both the higher and lower nouns were animate

[8]Cuetos and Mitchell (1988) relied largely on pragmatic disambiguation. See Section 2.2.1 of Fernández (2003) for useful discussion of these facts.

preferences. One such study, Fernández (2003), showed that offline judgments, in fact, diverged from self-paced reading behaviors. That is, while speakers of Spanish ultimately preferred interpretations in accordance with high attachment, their initial parsing behaviors registered a low attachment preference. In what follows, we explored whether neural models of language pattern in accordance with speakers interpretations. Ultimately, we find that models of Spanish do not.

4.4.1 Stimuli and Measures

For these experiments, we drew on stimuli from both Cuetos and Mitchell (1988) and Fernández (2003).[9] This amounted to 48 experimental items like:

(2) a. Andrew had dinner yesterday with the nephew of the teachers that was divorced.

b. Andrew had dinner yesterday with the nephews of the teacher that was divorced.

c. André cenó ayer con el sobrino de los maestros que estaba divorciado.

d. André cenó ayer con los sobrinos del maestro que estaba divorciado.

The underlined nominal above marks the attachment point of the relative clause (*that was divorced*).[10] That is verbal agreement in the relative clause picks out the attachment location. (2-a) and (2-c) exhibit HIGH attachment (i.e. *was* agrees

[9]The templates for the stimuli for the English and Spanish experiments are given in Appendix B.1.

[10]For bidirectional models, some modifications of the full relative clauses were necessitated. For example, some examples included conjoined verbs, which would both surface agreement features. Such models are typically evaluated by only masking one token. Thus, the inclusion of two agreeing verbs would either provide agreement information or violate the one token restriction.

with *nephew* and not *teachers*), while (2-b) and (2-d) exhibit LOW attachment. A full factorial design was used so that each stimulus had versions where the higher noun was plural and singular and the lower noun singular and plural. Nouns were checked for inclusion in the model vocabulary and substituted for synonyms when both the singular and plural form of the noun were not in the vocabulary of all models.

Recent work has demonstrated that targeted syntactic evaluations that utilize a small set of verb lemmas overestimates the systematicity of neural models of language (Newman et al., 2021). To address this potential limitation, we compared the total probability mass assigned to singular and plural verbs in the relative clause.[11] For English, verbs were taken from Newman et al. (2021), but restricted to the set of lemmas that had both singular and plural forms contained in the vocabulary of all the models to facilitate direct comparison.[12] This resulted in 1072 lemmas. For Spanish, verbs were taken from AnCora (Taulé et al., 2008) and Spanish GSD[13] and inflected using an inflector (mlconjug3; Diao, 2021) to generate pairs of singular and plural verbs for present, future, imperfect, and past tense. These are further filtered by the inclusion of both singular and plural forms being necessarily in the vocab of all Spanish models investigated. This yielded 384 verb pairs. For both languages, nouns were substituted for synonyms when both the singular and plural form were not in the vocabulary of all models.

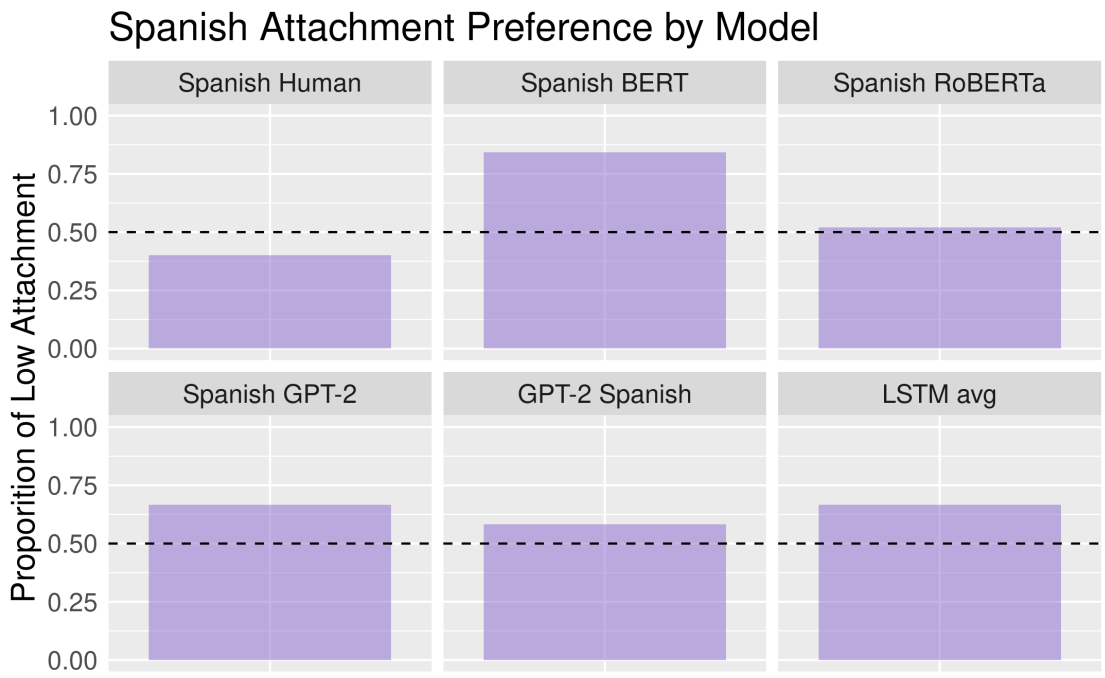

Figure 4.1: For Spanish, proportion of stimuli where low attachment was preferred for complex nouns for BERT, RoBERTa, Spanish GPT-2, GPT-2 Spanish, and by-item average of LSTMs (e.g., *the friends of the man who are...* over *the friends of the man who is...*). The dashed line depicts no preference. Stimuli and human results are from Cuetos and Mitchell (1988) and Fernández (2003).

4.4.2 Results

Results by model for Spanish are given in Figure 4.1 and for English in Figure 4.2. Statistical analyses[14] were conducted via linear-mixed effects models.[15] For

[11]For masked language models, we masked the target location (e.g., ... that [MASK] divorced). For autoregressive models we truncated the stimuli at *that*.

[12]We excluded lemmas which were subworded by the models with wordpiece vocabularies (e.g., GPT-2 XL).

[13]`https://github.com/UniversalDependencies/UD_Spanish-GSD`

[14]We used lme4 (version 1.1.23; Bates et al., 2015) and lmerTest (version 3.1.2; Kuznetsova et al., 2017) in R.

[15]We fit a model to predict the probability of the relative clause verb (which agrees with one of the nouns in the preceding complex NP) with an interaction between attachment height (high vs. low) and the number of the verb (singular or plural) and with random slopes by item for

English, the LSTMs, BERT, and GPT-2 XL all exhibited a general preference for agreeing with the lower noun (i.e. a LOW attachment preference). The LSTMs also had a preference for agreeing with plural nouns (but there was no interaction with attachment height). RoBERTa showed no preference for either attachment location. For Spanish, Spanish BERT and Spanish GPT-2 exhibited a general preference for agreeing with the lower noun (i.e. a LOW attachment preference). GPT-2 Spanish, RuPERTa, and the LSTMs both exhibited no preference for either attachment location. RuPERTa had a significant effect of verb number, preferring to agree with plural nouns.

4.4.3 Discussion

We found evidence that, despite cross-linguistic difference in human attachment preferences, neural models of English and Spanish favor the same attachment, LOW. A subset of neural models of Spanish and English had no preference (RoBERTa for English and GPT-2 Spanish, RuPERTa, and the LSTMs).

Depending on what aspects of human language processing we expect models to correlate with, these results may not be that surprising (see Chapter 2 for relevant discussion). Recall that Fernández (2003) found that despite cross-linguistic differences in ultimate attachment preferences (as measured by judgments from ambiguous sentences), online processing (as indexed by self-paced reading) was consistent between speakers of Spanish and English. That is, for Spanish speakers, initial processing behavior accords with a LOW preference (with greater reading

verb number and attachment height. For the LSTM models, random slopes by model for verb number and attachment height were also included. The statistical model for RoBERTa did not converge with any random effects, so a linear mixed effects model was fitted instead. The rest of the statistical models were maximal.

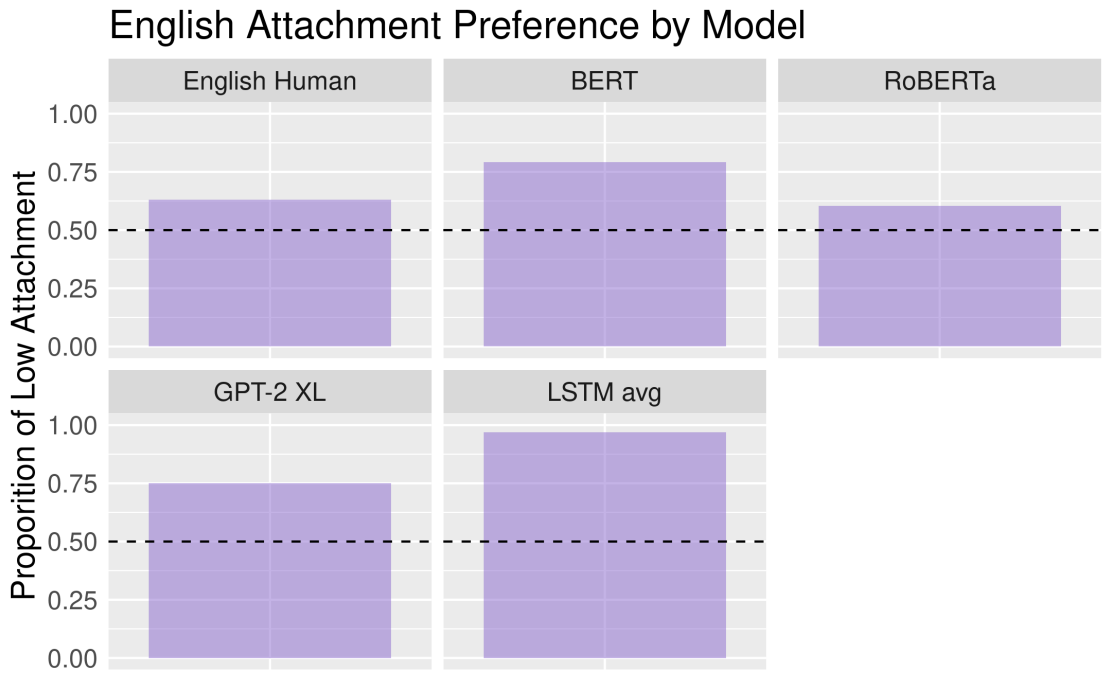

Figure 4.2: For English, proportion of stimuli where low attachment was preferred for complex nouns for BERT, RoBERTa, GPT-2 XL, and by-item average of LSTMs (e.g., *the friends of the man who are...* over *the friends of the man who is...*). The dashed line depicts no preference. Stimuli and human results are from Cuetos and Mitchell (1988) and Fernández (2003).

times with forced HIGH), which is presumably reversed in later processing.

It seems the present results point to the conclusion that neural models behave in accordance with human initial parsing, rather than ultimate interpretations. This may, in fact be desirable behavior. However, a large subset of research in natural language processing uses models like BERT for tasks requiring natural language comprehension. Here, we've demonstrated that BERT based models, along with others, fail to behave in accordance with the attachment preferences of Spanish speakers, and therefore, may be capturing a different set of biases (e.g., production biases) than those biases (e.g., comprehension biases) that certain tasks necessitate.

We return to these points in the general discussion (Section 4.8).

In what follows, we explored whether neural models capture some construction specific differences in attachment preferences in both English and Spanish, and then turn to a deeper examination of whether models capture other aspects of online attachment processing in English and Spanish. Our investigation of fine-grained attachment preferences demonstrates that neural models of both English and Spanish overemphasize a LOW attachment preference.

4.5 Fine-Grained Attachment Preferences in Neural Models

In the preceding section, we found that neural models of Spanish and English learned a LOW preference (or no preference) contrary to the cross-linguistic differences between speakers of these languages. Within psycholinguistics, there has been a number of studies attempting to reconcile the differences between Spanish and English speakers. As it pertains to this section, Gilboy et al. (1995) demonstrated that there is variation in attachment preferences depending on the specific constructions. In fact, English speakers have high attachment preferences for certain complex noun phrases.

Given that prior corpus work has suggested that fine-grained corpus frequencies track construction specific attachment preferences, at least for Dutch (Desmet et al., 2006), we might expect that neural models will capture this same variation. Ultimately, we find that neural models consistently obtain stronger LOW preferences than humans.

4.5.1 Stimuli and Measures

The stimuli for this experiment were taken from Gilboy et al. (1995) which explored predictions of Construal Theory (see Frazier and Clifton, 1996).[16] Construal Theory posits that relative clause attachment (or association within the theory) is modulated by properties of the possible attachment sites. Gilboy et al. (1995) isolated two such factors: (1) the referential status of the lower noun in complex noun phrases (cf. *the friend of the man* and *a cup of sugar* where sugar has no determiner) and (2) the argument status of the lower noun (cf. *the daughter of the colonel* and *the sauce with the steak*).

There were 3 broad experimental categories (i) Type A where the lower noun was a non-referential argument of the higher noun, (ii) Type B where the lower noun was a referential argument of the higher noun (these stimuli are closer to the ones used in other experiments probing attachment preferences; e.g., those in Section 4.4), and (iii) Type C where the lower noun was a referential non-argument of the higher noun.

They found that in offline judgments, Spanish and English speakers attachment preferences were largely similar (in contrast to work cited above), with a preference for high attachment most pronounced for Type A (e.g., *a sweater of wool*) and a low attachment preference most pronounced for Type C (e.g., *the house with a pool*). For Type B (e.g., *the side window of the plane*) both languages showed no marked preference. Additional manipulations were included (the use of indefinite or definite pronouns and the presence of adjectives). The by-item lower noun preferences were given in their appendix and collapse over these conditions, so we included them as

[16]The templates for the stimuli for the English and Spanish experiments are given in Appendix B.2.

additional forms to calculate the by-item average preferences for models.

In Gilboy et al. (1995), ambiguous stimuli were presented to participants and their attachment judgments elicited. This resulted in a low attachment preference score which was the proportion of responses that favored the lower attachment interpretation. In order to derive comparable preferences for computational models, we made use of a number manipulation. That is, for each item in the experiment, we constructed additional sentences where the higher and lower noun differed in number. Additionally, the further experimental manipulation in their paper (e.g., the use of indefinite vs. definite determiners) were included. This resulted in pairs of stimuli like:

(3) a. Andres picked up the sacks of sand that was brought from the construction site.

b. Andres picked up the sack of sands that was brought from the construction site.

Neural model preferences were categorically recorded as either (i) low if was was more probable in (3-a) than in (3-b), or (ii) high if was was more probable in (3-b) than in (3-a). The lower noun attachment preference was the proportion of cases where agreement with the lower noun was preferred to the higher noun. For the above, this included the addition of an adjective (brown), with BERT preferring low attachment 83% of the time and humans preferring low 56% of the time. Additionally, as with all these experiments, the number manipulation is carried by the nouns. Therefore nouns had to be checked for inclusion in the model vocabulary.[17]

[17]Some stimuli were skipped due to semantic substitution issues. Namely, A2 item 3 and A2

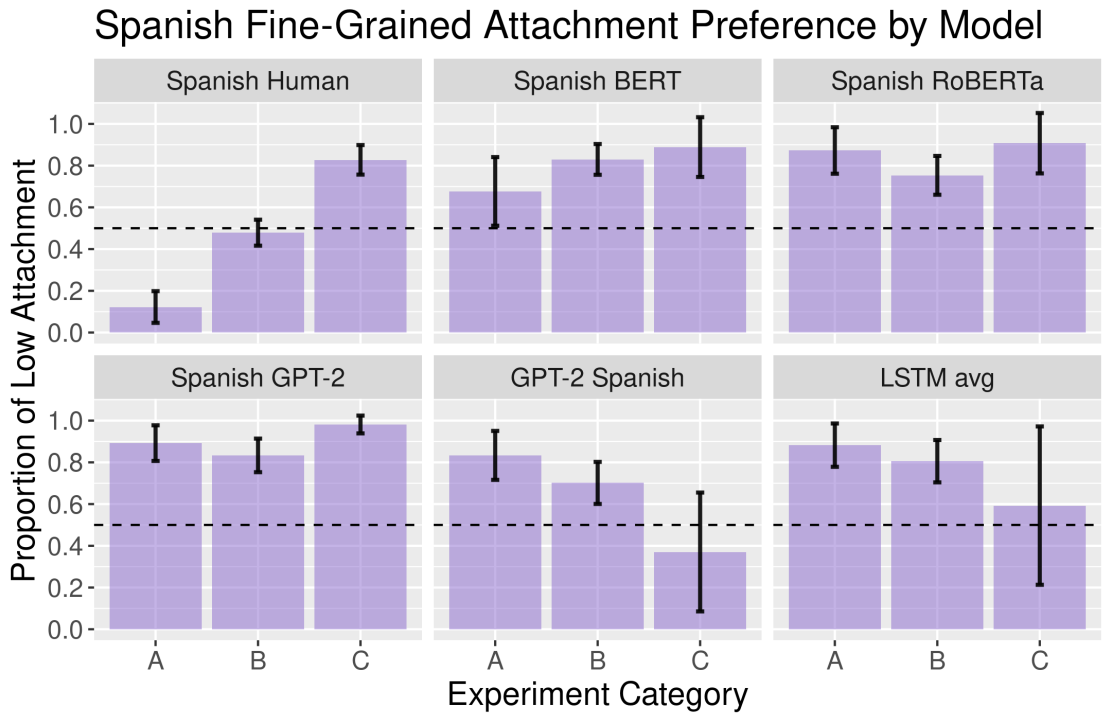

Figure 4.3: For Spanish, proportion of stimuli by experimental class that agreement with the lower noun was favored in a complex noun phrase for BERT, RoBERTa, Spanish GPT-2, GPT-2 Spanish, and by-item average of LSTMs (e.g., *man* in *the friends of the man*). Stimuli and human results are from Gilboy et al. (1995). Results are organized by three types of stimuli: Type A (non-referential lower noun; *a sweater of wool*), Type B (lower noun is a referential argument of the higher noun; *the side window of the plane*, and Type C (lower noun is a referential non-argument of the higher noun; *the house with a pool*).

4.5.2 Results

Results by model for Spanish are given in Figure 4.3 and for English in Figure 4.4. Statistical analyses were conducted via t-tests comparing the human and neural models' proportions of low attachment by category (i.e. across Type A, B, and C).[18] For English, all significant differences between humans and models skewed

item 6.

[18]To correct for multiple comparisons, an adjusted α of 0.007 was used.

towards a stronger low attachment preference for neural models as compared to humans. In looking at each model, BERT did not differ significantly from humans, keeping with the trend of a greater low attachment preference in Type C than Type A. RoBERTa had a significantly greater low attachment preference for Type B than humans, with no other differences. GPT-2 XL and the LSTM models all showed an overwhelming low attachment preference which was stronger than humans across all types. This is notable given that English participants exhibited a high attachment preference for Type A. Put another way, models that condition on both contexts were closer to the human preferences, capturing the gradation across the experimental conditions, while auto-regressive models, which only condition on the left context, exhibited an across the board preference for agreeing with the lower noun in all stimuli regardless of the experimental condition.

For Spanish, we once again see that neural models have a greater low attachment preference than humans. Spanish BERT, RuPERTa, Spanish GPT-2, GPT-2 Spanish and the LSTMs all showed significantly greater low attachment preferences for Types A and B than humans. This is particularly stark for Type A where the human high attachment preference was very strong (preferring low only 12% of the time), while all the neural models showed a preference for low attachment. Notably, the reverse held for Type C between humans and GPT-2 Spanish, where humans had a low attachment preference and GPT-2 Spanish had a high attachment preference.

4.5.3 Discussion

In this section, we found that neural models consistently learn stronger LOW preferences than humans. While confirming the results of Section 4.4, it extends the

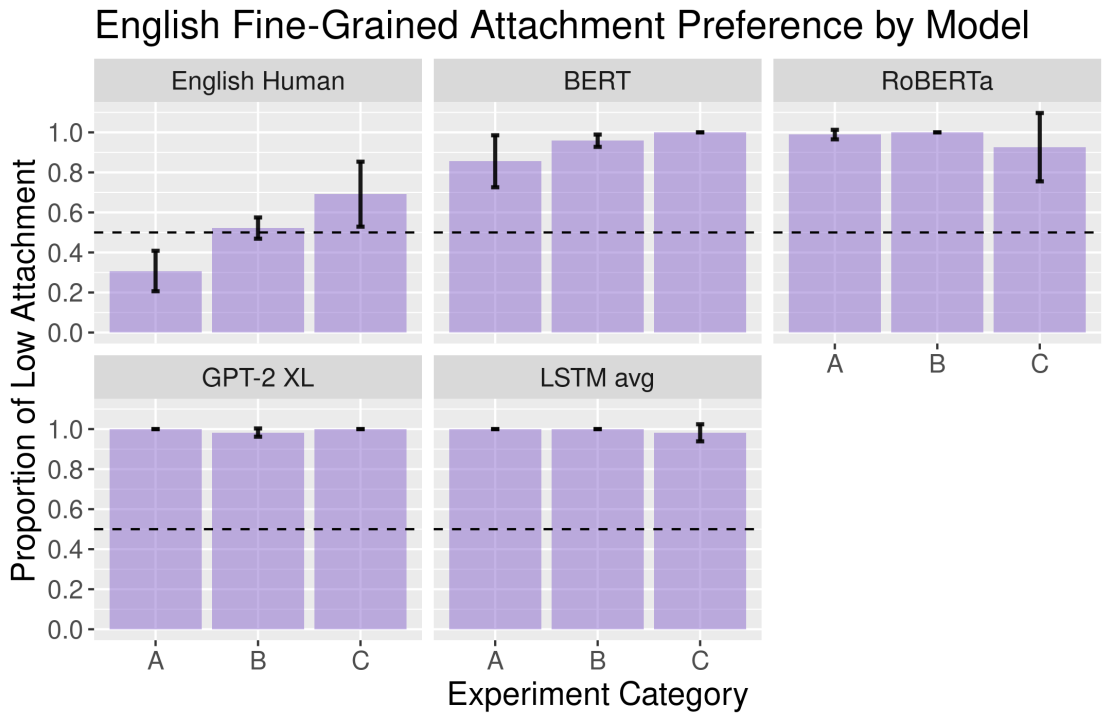

Figure 4.4: For English, proportion of stimuli by experimental class that agreement with the lower noun was favored in a complex noun phrase for BERT, RoBERTa, GPT-2 XL, and by-item average of LSTMs (e.g., *man* in *the friends of the man*). Stimuli and human results are from Gilboy et al. (1995). Results are organized by three types of stimuli: Type A (non-referential lower noun; *a sweater of wool*), Type B (lower noun is a referential argument of the higher noun; *the side window of the plane*, and Type C (lower noun is a referential non-argument of the higher noun; *the house with a pool*).

basic conclusions. Given that English speakers have a general LOW preference, we may have believed neural models were at least capturing the English-like pattern. However, Gilboy et al. (1995) demonstrated that, for certain constructions, English speakers also have an overwhelming high attachment preference. Neural models, on the other hand, do not capture this difference, favoring LOW regardless of condition. That is, neural models of English are not capturing the comprehension biases of English speakers.

Given that neural models have as their only training signal corpus frequencies, the present results cast doubt on whether fine-grained frequency counts engender the attested human interpretation preferences. In the final sections of this chapter, we explore additional cases of online parsing preferences (first in English and then in Spanish), finding further evidence that neural models resist an analysis where they fully capture any level of human linguistic processing.

4.6 Interaction between Attachment and Implicit Causality in English

In Chapter 3, we found evidence that implicit causality (IC) verb biases are learnable, at least to some extent, with transformer models. Under the standard paradigm of targeted syntactic evaluations (i.e. investigations of linguistic knowledge that focus on carefully controlled minimal pairs), we might conclude that current models know a human-like implicit causality bias. However, exploring linguistic phenomena in isolation obscures the underlying linguistic system of speakers.

In some linguistic theories (e.g., Optimality Theory; Prince and Smolensky, 2004) and psycholinguistic theories (e.g., MacWhinney et al., 1984), the interaction of linguistic processes (or constraints) is foundational to understanding human linguistic knowledge. We explored the interaction of IC verb bias and ambiguous relative clause attachment in English. We hypothesized that if a model did fully acquire a human-like IC verb bias and fit the online attachment preferences of English speakers, then it should be able to use this bias to modulate ambiguous relative clause attachment in line with humans. Looking ahead, we find that models fail to capture the interaction between attachment and implicit causality, suggesting

that their knowledge of implicit causality is not as robust as humans (see also Kementchedjhieva et al., 2021) or that their attachment preferences are not tracking online attachment preferences of humans.

4.6.1 Stimuli and Measures

We used the self-paced reading stimuli from Rohde et al. (2011), which consisted of 20 pairs of sentences.[19]

(4) a. Anna scolded the chef of the aristocrats who was/were routinely letting food go to waste.

b. Anna studied with the chef of the aristocrats who was/were routinely letting food go to waste.

The central manipulation in the self-paced reading study lies with whether the verb was an object-biased IC verb (*scolded*) or not (*studied with*). Human participants read sentences where the RC verb (e.g., *was* or *were*) either agreed with the higher noun (e.g., *chef*) or the lower noun (e.g., *aristocrats*). Rohde et al. (2011) reported decreased reading times for agreement with the higher noun when the verb was object-biased compared to when the verb was not object-biased. In other words, an object-biased IC verb facilitated attachment to the higher noun. In evaluating our models on these stimuli, we balanced them by number, so that the higher and lower noun were equally frequent as singular or plural in our test data. This resulted in 192 test sentences generated from 12 pairs.[20]

[19]The templates for the stimuli are given in Appendix B.3.

[20]We excluded pairs where either of the main verbs was not in the vocabulary of our LSTM LMs. There was one noun substitution, florist(s) with clerk(s). Additionally, we substituted

As with the above sections, we looked at the probabilities assigned by the neural models to target words. Following Rohde et al. (2011), we hypothesized that agreement with the higher nominal (e.g., *chef* in (4)) would be more probable after object-biased IC verbs than after the other verbs (e.g., agreement with *chef* should be more likely in (4-a) than in (4-b)). For autoregressive models, we evaluated the probability after the relative pronoun (i.e. *who* in (4)). For non-autoregressive models, we masked the relative clause verb:

(5) Anna scolded the chef of the aristocrats who MASK routinely letting food go to waste.[21]

For the above, BERT assigns a probability of 0.2% to singular and 68% to plural for (5), suggesting that BERT has a low attachment bias for this stimulus (contra the high attachment bias from humans).

4.6.2 Results

Recall, ambiguous relative clause attachment interacts with IC verb bias. For humans, object-biased IC verbs, in stimuli like *Anna scolded the chef of the aristocrats who...*, facilitated attachment to the higher noun (e.g., chef) in contrast to the general preference for attaching to the lower noun (e.g., aristocrats). We used the preferred number of the RC verb as a proxy for attachment location for each model. Results for each neural model (BERT, RoBERTa, LSTMs, GPT-2 XL) are given in

masculine names with *the man* and feminine names with *the woman*.

[21]Model specific mask tokens were used.

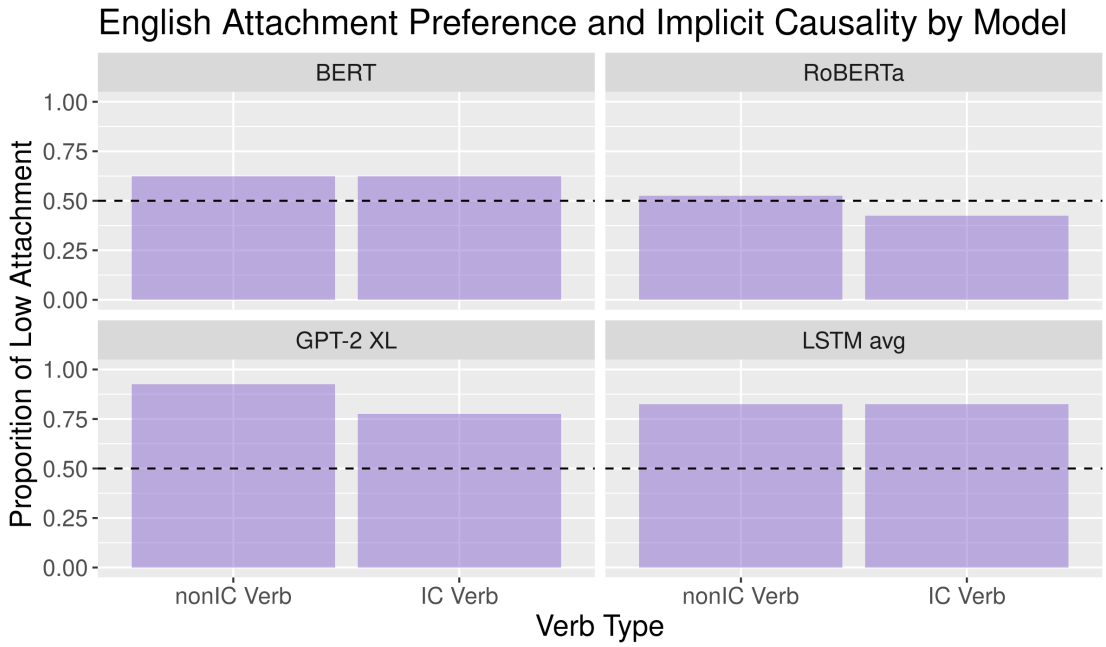

Figure 4.5: For English, proportion of stimuli where low attachment was preferred conditioned by IC verb bias for GPT-2 XL, BERT, RoBERTa, and by-item average of LSTMs. Stimuli are from Rohde et al. (2011) (e.g., *the woman scolded the chef of the aristocrats who verb...*).

Figure 4.5. Statistical analyses were conducted via linear-mixed effects models.[22]

The LSTM LMs, RoBERTa, and BERT had no significant interaction between IC verb bias and probability of attachment. GPT-2 XL did have a significant interaction between IC verb bias and attachment probability. However, IC did not influence high attachment, rather low attachment was more likely after non-IC verbs than object-biased IC verbs. GPT-2 XL, and the LSTM LMs all had a general preference for low attachment in line with the findings above. BERT and

[22]We fit a model to predict the probability of the pronoun with a three-way interaction between IC type (subject or object), pronoun antecedent (subject or object), and RC verb number (singular or plural) and with random slopes for items. For the LSTMs random slopes for item and model were included. Post-hoc t-tests were conducted to evaluate effects. The threshold for significance was set at 0.005.

RoBERTa have no attachment preference.

4.6.3 Discussion

In exploring the interaction between implicit causality and ambiguous relative clause attachment, we found that none of the investigated models captured an interaction in line with human studies. That is, despite a robust effect of IC verb bias in predicting pronouns for the transformer models, this IC knowledge failed to influence attachment preferences.

In contrast to our results, Davis and van Schijndel (2020b) showed syntactic predictions for LSTM LMs are influenced by some aspects of discourse structure. A simple explanation for these conflicting results may be that the LMs we examined here are unable to learn the syntactic operation of attachment, and thus no influence of discourse can surface. This would be in line with the preceding sections (see also Davis and van Schijndel, 2020c).

Ultimately, the converging results suggest that neural models of English fail to capture both the comprehension biases of English speakers (i.e. neural models always prefer low) and also the online processing behavior for attachment. In the final section, we return to Spanish and examine the relationship between gender agreement and attachment.

4.7 Gender Agreement and Attachment in Spanish

As discussed in Chapter 2 of Fernández (2003), the human experimental results suggest a mismatch between the processing of number agreement and gender agreement in Spanish. Namely, in online processing, attachment disambiguated by number favors LOW (e.g., Fernández, 2003), while attachment disambiguated by gender favors HIGH (e.g., Carreiras and Clifton, 1993).[23] Below, we explored whether neural models of Spanish realized different attachment preferences for gender agreement. Ultimately, we found that neural models did have different attachment preferences for gender agreement, namely HIGH, suggesting that this processing difference in humans may be contained in the linguistic signal (at least to some extent).

4.7.1 Stimuli and Measures

The stimuli for this experiment were taken from the grammatical gender disambiguation condition of Experiment 5 from Carreiras and Clifton (1993).[24] In their experiment, as with the number agreement experiments, a complex noun phrase was modified by a relative clause. However, instead of the higher and lower noun being distinguished by number, gender was contrasted. For example:

(6) a. La policía detuvo a la **hermana** del portero que estuvo **acusada** de hurto.

[23]This difference could be due to the fact that gender agreement is realized on adjectives (and not verbs) which occurs later in the relative clause, but we set aside an account of the human parsing differences and focus instead on the empirical findings (see Fernández, 2003, Chapter 2 for insightful discussion).

[24]The templates for the stimuli are given in Appendix B.4.

b. La policía detuvo al hermano de la **portera** que estuvo **acusada** de hurto.

In (6-a) and (6-b) the possible attachment locations are underlined, with the noun agreeing with the adjective (*acusada*) bold faced. In (6-a), the higher noun (*hermana*) is feminine while the lower noun (*portero*) is masculine. In (6-b), the reverse holds.[25] Therefore, the attachment location of the relative clause is distinguished by which noun agrees with the adjective in the relative clause. A total of 24 such stimuli were used in a self-paced reading experiment. Reading times were found to be significantly longer when the adjective agreed with the lower noun (e.g., portera in (6-b)).

In evaluating the neural models of Spanish with these stimuli, the adjective was either masked for masked language models, or the relative clause was truncated just prior to the adjective for auto-regressive models. For each stimulus, both masculine and feminine forms of each noun were used.[26] As with number, a distribution of agreement was measured, rather than just comparing the top predicted adjective. The set of Spanish adjectives evaluated were taken from Spanish AnCora (Taulé et al., 2008) and Spanish GSD.[27] To have pairs of masculine and feminine adjectives, handcrafted rules transformed inflected forms into masculine and feminine.

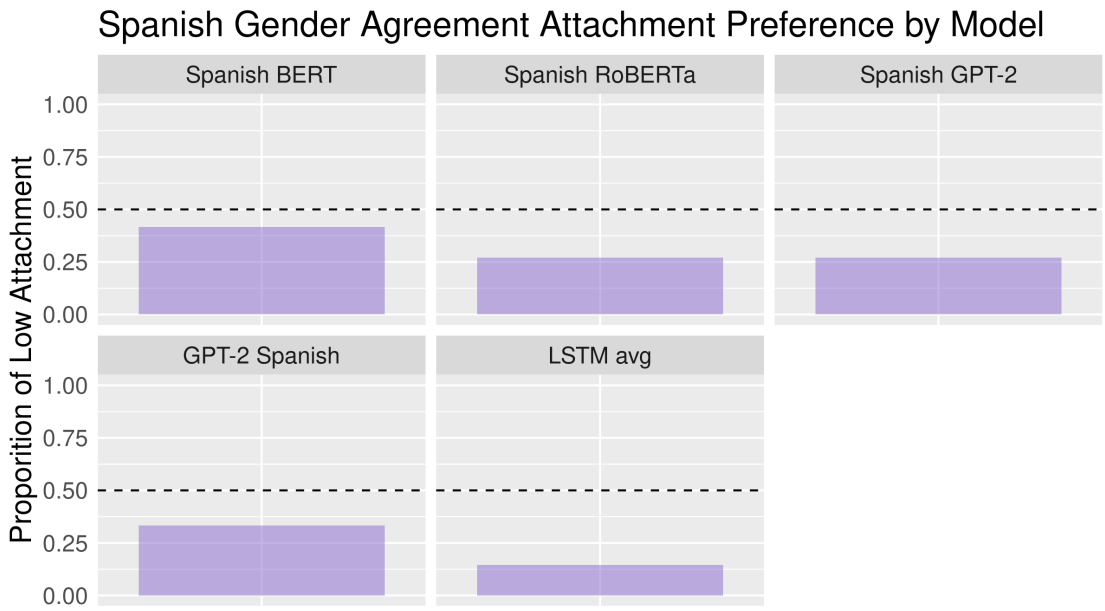

Figure 4.6: For Spanish, proportion of stimuli where RC adjective agreement with the lower noun in a complex noun was preferred for BERT, RoBERTa, Spanish GPT-2, GPT-2 Spanish, and by-item average of LSTMs (e.g., agreement with *man* in *the female friend of the man*). Stimuli are from Carreiras and Clifton (1993).

4.7.2 Results

Results by model for Spanish are given in Figure 4.6.[28] Statistical analyses were conducted via linear-mixed effects models.[29] Spanish GPT-2 and the LSTMs had

[25] Additionally, the determiners distinguish masculine from feminine nouns.

[26] Some nouns had to be substituted due to their absence in at least one models vocabulary.

[27] `https://github.com/UniversalDependencies/UD_Spanish-GSD`

[28] Note that in the figure, the probabilities are normalized by item and adjective gender. That is, for a context like *La policía detuvo NP1 NP2 que estuvo [MASK]*, the probability of feminine adjectives in the mask position was normalized by the total probability assigned to feminine adjectives when NP1 was feminine and when NP2 was feminine (i.e. $\frac{P(\text{acusada}|...\text{a la hermana del portero...})}{P(\text{accusada}|...\text{al hermano de la portera...})+P(\text{acusada}|...\text{a la hermana del portero...})}$).

[29] We fit a model to predict the probability of the relative clause adjective (which agrees in gender with one of the nouns in the preceding complex NP) with an interaction between attachment height (high vs. low) and the gender of the adjective (masculine vs. feminine) and with random slopes by item for adjective gender. For the LSTM models, random slopes by model for adjective gender were also included. The statistical models for BERT and GPT-2 Spanish did not converge,

a general preference for agreeing in gender with the higher noun (i.e. a HIGH attachment preference). Spanish BERT, RoBERTa, and GPT-2 Spanish had no effect of attachment location (i.e. neither a HIGH or a LOW attachment preference). RuPERTa, GPT-2 Spanish, and the LSTMs had a preference for feminine adjectives. In sum, rather than the low attachment preference noted in the above experiments, with gender agreement, the same models exhibited a high attachment preference.

4.7.3 Discussion

We found that, unlike the evidence from number agreement on the RC verb, neural models favored HIGH attachment when predicting adjective gender in the relative clause. That is, attachment preferences are modulated by which linguistic feature distinguishes the attachment. This is in line with the results from human studies, where Spanish speakers appear to favor LOW attachment initial parses when attachment is disambiguated by the RC verb number, and HIGH attachment when disambiguation is carried by gender on adjectives.

It seems, then, that neural models of Spanish pattern with the online parsing preferences of Spanish speakers, rather than with their later interpretation preferences. Results from English call for caution in generalizing these results, but nonetheless it may suggest a principled way of accounting for mismatches between neural models and humans. We leave to further work an examination of the interaction between gender and number, and a broader exploration of attachment preferences in related languages like Italian and French. This work does, however, make a prediction that in examining other languages with HIGH attachment preferences, models will only behave in accordance with human comprehension

so random slopes were removed.

biases when the immediate parsing behaviors of humans reflects the same bias. The present investigations, and its natural extensions, leave unresolved why such a mismatch exists (though see below for further discussion).

4.8 General Discussion

In this chapter, we explored the ability of neural models to prioritize multiple simultaneous valid interpretations in a human-like way (as in *John met the student of the teacher that was happy*). While both LOW attachment (i.e. *the teacher was happy*) and HIGH attachment (i.e. *the student was happy*) are equally semantically plausible without a disambiguating context, humans have interpretation preferences for one attachment over the other (e.g., English speakers prefer LOW attachment and Spanish speakers prefer HIGH attachment). Given the recent body of literature suggesting that neural models learn abstract syntactic representations, we tested the hypothesis that these models acquire human-like attachment preferences. We found that they do not.

We began by demonstrating that the general HIGH bias was not learned by neural models of Spanish. In fact, neural models of both English and Spanish demonstrated the same preference, LOW. Moreover, a fine-grained exploration of construction specific attachment preferences in English and Spanish showed that neural models consistently obtain stronger low attachment preferences than humans. Even in constructions where English speakers prefer high attachment, neural models of English favored low attachment. We take this to suggest a broader mismatch between neural models and human processing.

In the final sections, we more fully compared the online parsing preferences of

English and Spanish speakers to the behavior of neural models. Despite evidence in Chapter 3 that models learn implicit causality, and evidence suggesting models are capturing online parsing preferences of English, neural models fail to capture the interaction between these two processes. This suggests, in line with the conclusions throughout this dissertation, that neural models abstract less general linguistic knowledge than humans. For Spanish, we found that neural models capture an online processing difference in humans between number agreement and gender agreement, favoring HIGH attachment when disambiguated by gender.

In post-hoc analyses of the Spanish Wikipedia training corpus and the AnCora Spanish newswire corpus (Taulé et al., 2008), we find a consistent production bias towards LOW attachment among the RCs with unambiguous attachment. In Spanish Wikipedia, LOW attachment is 69% more frequent than HIGH attachment, and in Spanish newswire data, LOW attachment is 21% more frequent than HIGH attachment.[30] This distributional bias in favor of LOW attachment does not rule out a subsequent HIGH bias in the models. It has been established in the psycholinguistic literature that attachment is learned by humans as a general abstract feature of language (see Scheepers, 2003). In other words, human syntactic representations of attachment overlap, with prepositional attachment influencing relative clause attachment, etc. These relationships could coalesce during training and result in an attachment preference that differs from any one structure individually. However, it is clear that whatever attachment biases exist in the data are insufficient for neural models to learn a general human-like attachment preference in Spanish. This provides compelling evidence that standard training data itself may systematically lack aspects of syntax relevant to performing linguistic comprehension tasks.

[30] `https://github.com/UniversalDependencies/UD_Spanish-AnCora`

We suspect that there are deep systematic issues leading to this mismatch between the expected distribution of human attachment preferences and the actual distribution of attachment in the Spanish training corpus. Experimental findings from psycholinguistics suggest that this issue could follow from a more general mismatch between language production and language comprehension. In particular, Kehler and Rohde (2015, 2019) have provided empirical evidence that the production and comprehension of structures are guided by different biases in humans. Production is guided by syntactic and information structural considerations (e.g., topic), while comprehension is influenced by those considerations plus pragmatic and discourse factors (e.g., coherence relations). As such, the biases in language production are a proper subset of those of language comprehension. As it stands now, neural models are typically trained on production data (that is, the produced text in Wikipedia).[31] Thus, they will have access to only a subset of the biases needed to learn human-like attachment preferences. In its strongest form, this hypothesis suggests that no amount of production data (i.e. text) will ever be sufficient for these models to generalizably pattern like humans during comprehension tasks.

The mismatch between human interpretation biases and production biases suggested by this work invalidates the tacit assumption in much of the natural language processing literature that standard, production-based training data (e.g., web text) are representative of the linguistic biases needed for both natural language understanding and generation. There are phenomena, like agreement, that seem to have robust manifestations in a production signal, but the present work demonstrates that there are others, like attachment preferences, that do not. We speculate that the difference may lie in the inherent ambiguity in attachment, while agreement

[31]Some limited work has explored training models with human comprehension data with positive results (Klerke et al., 2016; Barrett et al., 2018).

explicitly disambiguates a relation between two syntactic units. This discrepancy is likely the reason that simply adding more data doesn't improve model quality (e.g., van Schijndel et al., 2019; Bisk et al., 2020).

Moreover, the possibility for ambiguity in syntactic attachment coupled with the pragmatic competence of speakers may account for the discrepancy between the biases in training data and those observed in human comprehension. Consider that speakers may know that comprehenders have certain preferences for disambiguating ambiguous attachment. In the case of Spanish discussed in this chapter, this would amount to speakers knowing that, when faced with an instance of ambiguous relative clause attachment, comprehenders will favor the interpretation following from high attachment. They may use this information to produce more cases of ambiguous relative clause attachment when the intended meaning follows from high attachment (comprehenders will likely infer that after all). Ultimately, the resultant outputs of speakers will have more cases of unambiguous low attachment (i.e. the output for which comprehenders do not have a bias to interpret) than unambiguous high attachment, leading models to learn low attachment as the more general form. Future work needs to be done to understand more fully what biases are present in the data and learned by language models, but, nonetheless, this direction seems promising.

Although this chapter raises questions about mismatches between human syntactic knowledge and the linguistic representations acquired by neural language models, it also shows that researchers can fruitfully use sentences with multiple interpretations to probe the linguistic representations acquired by those models. Evaluations have largely focused on cases of unambiguous grammaticality (i.e. ungrammatical vs. grammatical). By using stimuli with multiple simultaneous valid

interpretations, we found that evaluating models on single-interpretation sentences overestimates their ability to comprehend abstract syntax. Moreover, we posit that further explorations of mismatches between online parsing preferences and subsequent interpretations in humans, will point to further mismatches between neural models and humans and clarify the boundary between the two.

CHAPTER 5

PRINCIPLE B AND COREFERENCE

5.1 Introduction

The preceding chapters dealt primarily with ambiguity resolution, where multiple interpretations are grammatically licensed, but where humans have a preference for one particular form (e.g., preferring continuations like *is tall* to *are tall* for contexts like *the friends of the man who...*).[1] The failure of a neural model of a given language to mimic human preferences can, nonetheless, yield a grammatically acceptable form, just one that is odd or dispreferred.[2] The present chapter explores, instead, cases where some alternative structures (or more aptly behaviors) are ungrammatical. Namely, we explore the role of syntactic constraints on limiting possible coindexation between pronouns and antecedents.

In Chomsky (1981), three principles constraining the distribution of pronouns and anaphora (and their respective interpretations relative to possible antecedents) were proposed:

(1) **Binding Principles**

PRINCIPLE A: An anaphor is bound in its governing category

PRINCIPLE B: A pronominal is free in its governing category

PRINCIPLE C: An R-expression is free

[1] Code for replicating the experiments, figures, and statistical models in this chapter can be found on Github at `https://github.com/forrestdavis/Dissertation/tree/main/Binding`. Templates for the stimuli are provided in Appendix C for ease of reference.

[2] This is not to say that neural models are arriving at grammatical structures (i.e. building something like a linguistic parse). Rather, it is to say that the observed behavior is consistent with another possible grammatical structure.

Roughly, Principle A excludes examples like *John thinks that Keisha likes himself* from meaning JOHN THINKS THAT KEISHA LIKES JOHN (i.e. John is coindexed with *himself*). Principle B excludes examples like *John hates him* from meaning JOHN HATES JOHN (i.e. John is coindexed with him). Finally, Principle C excludes *He hates John* from meaning JOHN HATES JOHN (i.e. he is coindexed with John). While coreference more generally is modulated by discourse and pragmatic structure, binding principles are a structural relation mediated by c-command. In other words, the possible relations between anaphora and pronouns and antecedents is mediated by a structural (syntactic) relationship. While the specific binding conditions have been refined within syntactic theory (e.g., Reinhart and Reuland, 1993), we focus on some empirical consequences of Principle B that hold regardless of the underlying theoretical implementation.

Consider the following sentence:

(2) Bill told Clark that Robert had deceived him.

Despite *him* agreeing in gender with *Bill*, *Clark*, and *Robert*, not all three noun phrases are possible antecedents. Principle B stipulates that *him* can not corefer with *Robert*. Therefore only two interpretations are possible: ROBERT DECEIVED CLARK or ROBERT DECEIVED BILL.

In what follows, the extent to which neural models of English constrain their predictions in accordance with Principle B is explored. We begin by evaluating neural models' predictions of pronouns, following experimental work from humans. While models appear to entertain more antecedents for pronouns than humans, we find behavior in adherence to Principle B. However, as with the preceding

chapters, models fail to capture more complex interactions between Principle B and other linguistic processes. Ultimately, these results suggest that the parsing behaviors exhibited by humans do not always follow directly from experience with linguistic data. Moreover, fundamental structural constraints on human linguistic processing, while seemingly recoverable from linguistic data to some extent, are not fully abstracted in neural models.

5.2 Background

As noted above and in Chapter 3, the referent of ambiguous pronouns is constrained by structural considerations and modulated by semantics or discourse relations like implicit causality (or general biases like a preference to agree with the subject). An ongoing question within psycholinguistics is the time course of these processes: when encountering an ambiguous pronoun which antecedents are initially considered?

Concretely, for (2) we may focus on two considerations: (i) gender agreement between the pronoun and possible antecedent, and (ii) the structural restrictions imposed by Principle B. Within psycholinguistics, these are often thought of as constraints which drive measurable behaviors in humans (e.g., violating Principle B by linking a pronoun to a structurally illict antecedent results in a processing cost). It is then natural to ask whether constraints are evaluated simultaneously (e.g., MacDonald, 1994) or in some specific order (e.g., agreement checking operates before Principle B).

A number of studies have attempted to determine whether structural constraints operate immediately. Often this is operationalized by assuming that when encountering a pronoun, a set of possible candidate antecedents is returned to be possibly

linked with the pronoun. Consider the following sentence from Chow et al. (2014):

(3) Bill explained to Mary that Peter had deceived him.

The maximal candidate set, excluding the possibility of extra-sentential referents, is {Bill, Mary, Peter}. If agreement was an initial filter (prior to Principle B), the candidate set would be restricted to {Bill, Peter}. That is, *him* might pick out *Peter* which is ultimately ungrammatical. If instead, agreement and Principle B operate simultaneously, the candidate set would be {Bill}. Assuming that successfully returning a possible antecedent leads to comparable early behavior measures (i.e. violations of Principle B for coreferencing *him* with *Peter* are realized in later processing), these two hypotheses yield different predictions for sets of stimuli like:

(4) a. Bob thought Jim hated him
b. Bob thought Sue hated him
c. Mary thought Jim hated him
d. Mary thought Sue hated him

If agreement restricted the candidate set first, then we expect (4-a) – (4-c) to pattern together (as they all return at least one possible antecedent) to the exclusion of (4-d) (which returns no candidate sets). If Principle B and agreement operate in tandem, we expect (4-a) and (4-b) to pattern together (they both yield one possible antecedent *Bob*) and (4-c) and (4-d) to pattern together (they both yield no possible antecedents). Put another way, whether the embedded subject (*Jim* or *Sue*) influences behavior at *him* is determinate of Principle B operating earlier or later in processing.

A number of works have found that structural constraints immediately constrain the set of possible antecedents (e.g., Clifton et al., 1997; Sturt, 2003; Chow et al., 2014; Kush and Phillips, 2014; Kush and Dillon, 2021). That is, finding that (4-a) and (4-b) pattern together and (4-c) and (4-d) pattern together. However, other work has suggested that grammatically illicit antecedents can in fact have measurable effects (e.g., Badecker and Straub, 2002; Kennison, 2003). In other words the initial set may contain ungrammatical antecedents as well. It may well be that there are task specific effects from the means of measuring behavior (e.g., reading times vs. cross-modal priming) that drive different candidate sets (as discuseed in Nicol and Swinney, 2003), or that different measures capture different time points in processing, with later stages of processing potentially adding grammatically illicit candidates (Sturt, 2003). Ultimately, the plurality of the evidence suggests that Principle B has immediate effects on processing with possible repair mechanisms following later.

Turning to the question of computationally modeling the human parser, the dominant framework adopted by researchers studying the influence of binding conditions on parsing is cue-based retrieval (see for example Lewis and Vasishth, 2005). The parser under this framework relies on content addressable memory retrieval, which many works have suggested has no straightforward way of implementing the c-command relation necessary for articulating binding conditions (e.g., Kush, 2013; Dillon et al., 2013; Kush and Phillips, 2014; Kush et al., 2015). The origin of these hard structural restrictions of pronominal constructions remains unclear (for futher discussion on this and related topics see Kush, 2013). Given that recent work has connect autoregressive transformer models (namely, GPT-2 XL) to cue-based retrieval, we may expect difficulty for such models to capture Principle B (for relations between GPT-2 XL and cue-based retrieval see Ryu and Lewis, 2021).

Nonetheless, existing models have been claimed to capture (at least superficial) aspects of Principle A (see for example Warstadt et al., 2020a). While focusing their analysis on Principle A, the contrast Warstadt et al. (2020a) explored implicates Principle B by comparing acceptable coreference with *him* to unacceptable coreference with *himself*.[3] Additional work, has suggested that neural models (namely, BERT, TransformerXL, and some LSTMs) learn conditions on reflexive anaphora, again in line with Principle A (Hu et al., 2020a). Taken together, current evaluations of neural models of language suggest that binding conditions may be acquirable just from text.

The present study straightforwardly extends existing studies of neural models to Principle B. While we cannot assess whether neural models truly "interpret" the pronoun as coindexing with certain antecedents, we can compare the behavioral differences in neural models with minimally different stimuli. In fact, human online sentence comprehension studies are similarly limited. Online reading times are taken as a proxy for the consideration of certain antecedents, as we can not directly measure the content retrieved in reading a pronoun. Along these lines, we begin by replicating the experiments in Chow et al. (2014) with neural models of English. Then we proceed to evaluating the descriptive generalizations for model behavior in an adaption of Nicol and Swinney (1989) which affords more possible antecedents. Finally, we turn to the relation between forward prediction of pronominal referents (i.e. the processing of cataphora as in *While he was eating, Bill laughed.*) and Principle B (exemplified by Kush and Dillon, 2021). To preempt our results, we found that Principle B is not fully learned by any neural models, instead only some

[3]Warstadt et al. (2020a) state that because "coindexation cannot be annotated in BLiMP, Principles B and C are not illustrated" (p. 381). Nonetheless by relying on "him" in contrast to "himself", they are using a proxy for coindexation. This is, in some sense, inescapable given that Principle A and B are often thought as complementary, though there are deeper theoretical distinctions between the two (see Reinhart and Reuland, 1993).

isolated behaviors in accordance with it are attested.

5.3 Neural Models and Measures

We analyzed both long short-term memory networks (LSTMs; Hochreiter and Schmidhuber, 1997) and transformer models throughout the present chapter. For English, we used the 25 LSTM models trained on Wikitext-103 (Merity et al., 2016) that were detailed in Chapter 3. For transformers, we evaluated BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), TransformerXL (Dai et al., 2019), and GPT-2 XL Radford et al. (2019) via HuggingFace (Wolf et al., 2020).

Rather than assessing the probability assigned to pronoun as in Chapter 3, we used surprisal (Hale, 2001; Levy, 2008). This allowed us to compare reading times from humans and our model predictions (see Section 5.8 for the utility of this comparison), where appropriate (i.e. only for autoregressive models), following van Schijndel and Linzen (2018a); Wilcox et al. (2020a). Surprisal is defined as:

$$-\log \text{P}(\text{word}|\text{context})$$

For autoregressive models (e.g., GPT-2 XL) the context is the preceding words (e.g., for *the dog*, the surprisal for *dog* is -log P(dog| the)). For non-autoregressive models (e.g., BERT) the context is the surrounding non-masked tokens (e.g., for *the dog is happy*, the surprisal of *dog* is -log P([MASK] = dog | the [MASK] is happy)).[4]

[4]It is worth repeating that direct comparisons between reading times and the surprisal assigned to a word for non-autoregressive models is not straightforwardly possible as the model predictions can be conditioned on material not accessible to human readers. It is included in the present chapter only for ease of comparisons between models. Even there, though, caution is advised.

For ease of interpretation, we further calculated a gender mismatch effect (GMME) using the surprisal values. While our statistical tests were conducted on the surprisal values, GMMEs are used in the figures of the chapter. In human experiments, GMMEs index the increased cost in processing incurred when encountering a pronoun (or an antecedent) with an unexpected gender (e.g., van Gompel and Liversedge, 2003; Reali et al., 2015; Kush and Dillon, 2021). Thus GMMEs are a means of measuring human predictions by providing evidence for mismatches between expectations and reality. For neural models, we calculated GMMEs for both pronoun prediction and antecedent prediction (when the pronoun is cataphoric and, thus, precedes its antecedent).

For predictions about upcoming pronouns, consider:

(5) a. Fred thought Kathy hated him
    b. Mike thought Kevin hated him

To calculate the GMME for stimuli like (5), we took the difference between the surprisal for *him* in (5-a) and the surprisal for *him* in (5-b). A positive GMME would suggest that the model was more surprised when the embedded subject mismatched in gender with the pronoun; in other words, the gender of the embedded subject influenced the surprisal of the pronoun. In this case, comparing the GMME for *him* and *his* is informative about the status of Principle B in neural models. Humans have been shown to exhibit no GMME dependent on the embedded subject with *him*, because Principle B blocks co-indexation between these positions. For *his*, however, co-indexation is possible, and a GMME is obtained (see Chow et al.,

Ultimately, other works (and the above chapters) have evaluated such models, so they are included presently for continuity with those works.

2014).

For predictions about upcoming antecedents after cataphoric pronouns, consider:

(6) a. While he was at work, Fred ate food.

b. While he was at work, Keisha ate food.

For stimuli like (6), we calculated a GMME by taking the difference in surprisal of *Keisha* in (6-b) and the surprisal of *Fred* in (6-a). A positive GMME would indicate that the neural model was more surprised when the subject mismatched with the gender of the cataphoric subject pronoun.

5.4 Principle B as a Constraint on Accessibility: 2 NPs

In assessing the grammatical (and/or semantic) restrictions on coreference, a common framing is to ask which antecedent noun phrases are accessible at the point the pronoun is encountered (see Section 5.2 for further discussion). Consider, for example, the following sentence:

(7) Bill thought Sue hated him.

At the pronoun *him*, two noun phrases have been encountered, *Bill* and *Sue*. The question then is in processing *him* which of these noun phrases is considered. For humans, a growing body of literature has suggested that only grammatical licit positions are considered (e.g., Clifton et al., 1997; Sturt, 2003; Chow et al., 2014; Kush and Phillips, 2014; Kush and Dillon, 2021). In the case of (7), Principle B

and gender agreement block coreference between *Sue* and *him*, thus only *Bill* is retrieved.

In what follows, we explored the degree to which neural models of English pattern like humans in which noun phrases influence the pronoun. To foreshadow the results, we found that a subset of neural models (BERT, TransformerXL, and the LSTMs) patterned like humans, in that only the gender of grammatically licit antecedents influenced the surprisal of the pronoun. Another subset of the neural models (GPT-2 XL and RoBERTa) showed influences of the gender of both grammatically licit and illicit antecedents, though gender agreement with illicit antecedents increased the surprisal of pronouns (in accordance with a cost to violating Principle B).

5.4.1 Stimuli

The stimuli in this section are drawn from the Common Nouns condition from Experiment 1 in Chow et al. (2014).[5] There were 60 sets of stimuli contrasting whether the main clause subject (in bold) gender matched the pronoun and whether the embedded clause subject (in italics) gender matched the pronoun (i.e. a 2 X 2 design). An example set is given below (item 53 from Chow et al. (2014)).

(8) a. **Martin** dreamed that the *wizard* would poison him surreptitiously on the night of the full moon. (Match, Match)

b. **Martin** dreamed that the *witch* would poison him surreptitiously on the night of the full moon. (Match, Mismatch)

[5]The stimuli templates are given in Appendix C.1.

c. **Brenda** dreamed that the *wizard* would poison him surreptitiously on the night of the full moon. (Mismatch, Match)

d. **Brenda** dreamed that the *witch* would poison him surreptitiously on the night of the full moon. (Mismatch, Mismatch)

Due to neural model vocabulary restrictions, all proper names were replaced with *the man* or *the woman*. Only the masculine pronoun was considered, since the feminine pronoun *her* is ambiguous between an object pronoun, which will be subject to Principle B restrictions, and a possessive pronoun (e.g., *her friend*) which is not similarly restricted. For non-autoregressive models like BERT the full sentence was given with the pronoun (*him*) replaced with the relevant mask token.

Later experiments in Chow et al. (2014) contrasted *him* and *his*. For our experiments, we added an additional condition for *his* for the non-autoregressive models (no additional stimuli were needed for autoregressive models where we can easily compare the probability of *his* and *him* given the same context). These additional stimuli were adapted from their Experiment 3. For example, (8) had additional stimuli like *Martin dreamed that the wizard would poison his lover surreptitiously on the night of the full moon.*

Recall, we are interested in which stimuli in sets like (8) pattern together. For humans (8-a) and (8-b) pattern together (as do (8-c) and (8-d)). Reading times at the pronoun *him* pattern with the gender of the matrix subject (*Martin* or *Brenda*), with no reading time difference condition on the gender of the embedded subject. Chow et al. (2014) took this as evidence that Principle B immediately restricts the possible antecedents of *him*.

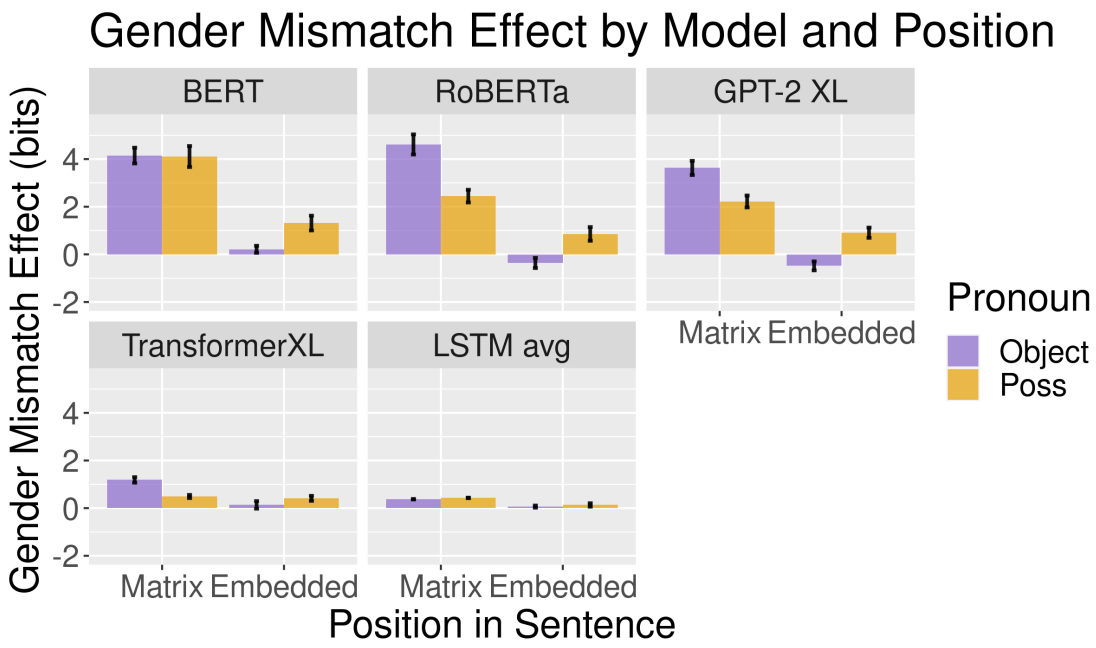

Figure 5.1: GMME for object pronoun (*him*) and possessive pronoun (*his*) by whether i) the matrix subject, or ii) the embedded subject agrees in gender (e.g., *the (man/woman) thought the (boy/girl) hated him*). A positive GMME means the pronoun gender was predicted to agree with the antecedent. A negative GMME means the pronoun gender was predicted to disagree with the antecedent. Error bars are 95% confidence intervals. Stimuli adapted from Chow et al. (2014).

5.4.2 Results

Results by model for English are given in Figure 5.1. Statistical analyses[6] were conducted via linear-mixed effects models.[7] Additionally pairwise t-tests were

[6]We used lme4 (version 1.1.23; Bates et al., 2015) and lmerTest (version 3.1.2; Kuznetsova et al., 2017) in R.

[7]We fit separate models for each pronoun (*him* or *his*) to predict the surprisal of the pronoun with an interaction between matrix subject gender (cf. *the man thought the woman hated him* and *the woman thought the girl hated him*) and the embedded subject gender (cf. *the man thought the boy hated him* and *the man thought the girl hated him*) with random slopes by item for the matrix and embedded subjects genders'. For the LSTM models, random slopes by model for matrix and embedded subjects genders' were included. Random effects were removed if the statistical model failed to converge.

conducted to assess differences between the experimental conditions.[8]

As discussed above, we interpreted our results relative to which stimuli groups differ statistically. That is, do stimuli with matrix subjects that agree with the pronoun pattern together to the exclusion of the other stimuli? We first present results for the pronoun *him* which interacts with Principle B, before giving results for the pronoun *his* which does not. Recall, that the human results from Chow et al. (2014) suggested two groupings of stimuli: one group where the matrix subject agrees with the pronoun (e.g., *the man thought the woman hated him*) and another group where the matrix subject does not agree (e.g., *the woman thought the man hated him*). The results from the neural models fell into two bins. One bin aligned with the human results and one did not.

For BERT, TransformerXL, and the LSTMs, only the gender of the matrix subject influenced the surprisal of *him* in line with only grammatically licit positions being considered for coreference (mirroring the human results in Chow et al. (2014)). We could organize this in a "ranking" of conditions like {MM, MF} < {FM, FF}, where < means less surprising and brackets group equally surprising conditions. For RoBERTa and GPT-2 XL, both the gender of the matrix subject and the gender of the embedded subject influenced the surprisal of *him*. In terms of experimental conditions, the stimuli where the matrix subject agrees with the pronoun are the least surprising, followed by the condition where both the matrix and the embedded subjects disagree with the gender of the pronoun, and lastly the condition where the matrix subject mismatches with the pronoun and the embedded subject agrees (i.e. {MM, MF} < FF < FM). These groupings differ from humans and suggest a ungrammatical gender match effect, as in Badecker and Straub (2002) and Kennison (2003) though with qualifications (see below for discussion).

[8]To correct for multiple comparisons an adjusted α of 0.0008 was used to determine significance.

For the pronoun *his*, all models exhibited the same pattern; the gender of both the matrix subject and the embedded subject influenced the surprisal of the pronoun. In contrast to *him*, we have a surprisal "ranking" of {MM, MF} < FM < FF.

5.4.3 Discussion

In the above experiments, we found evidence for two classes of neural model behavior. One appeared to follow that of the human behavior in Chow et al. (2014): only the gender of the grammatical licit antecedent (in this case, the matrix noun phrase), influenced the surprisal of the object pronoun (*him*). The other class of neural model behavior included some influence of the grammatically illicit antecedent (in this case, the embedded noun phrase). This pattern appears to fit with a descriptive generalization that the pronoun is equally surprising when the matrix subject agrees in gender (i.e. the gender of the other noun phrase showed no effect), but when the matrix subject did not agree, there was a cost associated with agreeing with the ungrammatical antecedent.

As alluded to above, this pattern of increased surprisal for ungrammatical agreement mimics some findings in the human processing literature. Badecker and Straub (2002) found evidence that humans consider both grammatical and ungrammatical antecedents in pronouns suggesting that Principle B is evaluated in parallel with other constraints, rather than acting as an initial filter. Their results showed an increased processing cost when both the grammatical and ungrammatical antecedent agreed in gender (i.e. the MM condition was read more slowly than MF).

Within a parallel constraint model, we might assume there is a constraint which targets agreement between a pronoun and an ungrammatical antecedent. Such a constraint would predict an effect on surprisal for both the MM and the FM conditions, and would accord with the findings in Badecker and Straub (2002). However, the behavior of neural models seems to conform to a conditional constraint, which targets agreement between a pronoun and an ungrammatical antecedent only when there is no agreeing grammatical antecedent. An effect like this has been documented for humans, namely in Sturt (2003) and Kennison (2003).

In both these works, it is assumed that there are two stages of processing. For example, in the general discussion of Kennison (2003) it is stated that:

> When a good match is found, the process of antecedent search is terminated, eliminating the possibility that structurally unavailable candidate antecedents will influence processing. When a possible match is not found or a possible match is evaluated and found to be weak, the process of antecedent search continues. During this time, a structurally unavailable antecedent may influence processing. (Kennison, 2003, p. 351)

Under a conception of antecedent search like the above, a process is engaged by the human parser when encountering a pronoun that unfolds in time, subject to earlier or later termination depending on the particular stimulus properties. It is difficult to imagine, however, that neural models of language have an embedded computation that works in this sequential manner, first looking for grammatically available antecedents and then searching a broader set of antecedents. Likely there must be another explanation for the attested neural model behavior.

A tentative proposal is that neural models highly weight agreement with the subject with an additional threshold of activation that saturates (i.e. after some degree of activation, any influence of agreement with ungrammatical antecedents fails to influence the surprisal at the pronoun). Then, the difference between the classes of neural model behaviors is driven by an additional Principle B like constraint that can increase surprisal for ungrammatical antecedents. The other models (which show only an influence of the matrix subject) merely track the properties of the matrix subject.

The results with the pronoun *his*, which is not subject to Principle B violations, partially confirmed this proposal. We found that the surprisal of *his* is influenced by both the matrix and the embedded subject. However, agreement with the subject was preferred (i.e. MM or MF was preferred to FM). Models have a general preference for agreeing with the matrix subject. Additionally, the difference in behavior between *his* and *him* demonstrates that neural model behaviors for *him* are not (entirely) general behaviors for coreference, and instead are conditioned on that specific pronoun. That is for neural models where the surprisal of *him* was only influenced by the matrix subject, the models did ignore the ungrammatical antecedents in line with Principle B (and contrary to their biases for *his*). This might suggest some knowledge of Principle B, or at least a greater subject preference for object pronouns. In what follows, we try to tease apart general subject preferences from behavior mimicking that of Principle B.

5.5 Principle B as a Constraint on Accessibility: 3 NPs

The above experiment involved two noun phrases (e.g., NP1 *thought* NP2 *hated him*). While measurable effects conditioned on the gender of the embedded noun point towards Principle B not operating as an initial filter on antecedents (i.e. coreference with the lower noun indicated by agreement is ungrammatical, and therefore should not occur), measurable effects conditioned on the matrix subject are not fully discriminative. That is, a simple heuristic could account for only agreeing with the matrix subject, namely, agreeing with the first noun (or the subject more narrowly).

In fact, the human results in Chow et al. (2014) seem to follow this simple heuristic. The pronoun *his*, should be able to agree with either noun, however, as with *him*, only the matrix subject has a demonstrable influence on surprisal. Neural models showed a difference between *his* and *him*, ruling out a simple subject preference for *his*. However, this does not rule out a simple heuristic tied to the pronoun *him*. The following experiment seeks to address this inferential shortcoming.

Inspired by Nicol and Swinney (1989), we expanded the number of possible antecedents to three, where two were grammatically possible antecedents and one was ungrammatical due to Principle B. We found more varied behavior for the neural models with three noun phrases than with two. Some neural model behaviors reduced to tracking masculine gender (LSTMs and TransformerXL). Others showed a penalty for gender agreement between the pronoun and ungrammatical antecedents (BERT, RoBERTa, and GPT-2 XL).

5.5.1 Stimuli

The present study adds an additional noun phrase in order to tease apart an adherence to Principle B and a simple subject preference. The stimuli are adapted from Experiment 2.1 in Nicol (1988) (discussed also in Nicol and Swinney, 1989).[9] In the study, cross-modal priming was used to determine which antecedents were reactivated by the pronoun in stimuli like:

(9) The landlord told the janitor that the fireman with the gas-mask would protect him if it became necessary.[10]

Nicol (1988) found that *him* only reactivated *landlord* and *janitor*. In other words, *him* reactivated both syntactically available noun phrases, blocking *fireman* in accordance with Principle B. The 24 stimuli used in that experiment, were adapted to form sets of stimuli that included all combinations of gender for the three noun phrases (e.g., NP1[m/f] told NP2[m/f] that the NP3[m/f] would protect *him*...). As detailed in the prior section, only *him* was considered, with right contexts provided for non-autoregressive models.

If the neural models consider only antecedents in accordance with Principle B, then the surprisal of the pronoun should be conditioned only on NP1 and NP2. If instead, the first noun heuristic (or subject preference) describes the neural model behavior correctly, then only NP1 should influence the surprisal of the pronoun. Critically, an influence of NP3 is not observed in humans and so any NP3 influence indicates non-human-like behavior (see Nicol, 1988).

[9]The stimuli templates are given in Appendix C.2.
[10]This is from Nicol (1988, p.65).

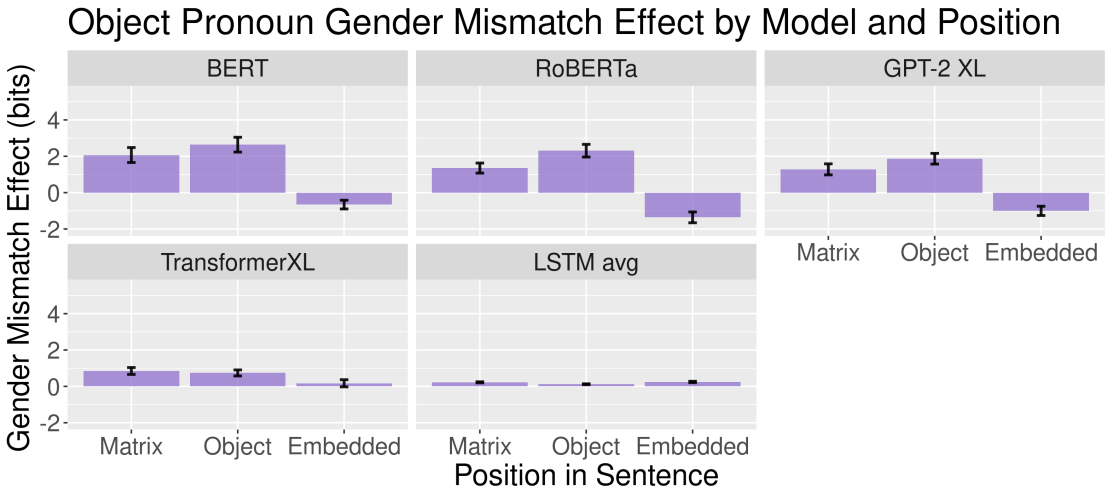

Figure 5.2: GMME for object pronoun (*him*) by whether i) the matrix subject, ii) the matrix object, or iii) the embedded subject agrees in gender (e.g., *the (man/woman) told the (prince/princess) that the (boy/girl) hated him*). A positive GMME means the pronoun gender was predicted to agree with the antecedent. A negative GMME means the pronoun gender was predicted to disagree with the antecedent. Error bars are 95% confidence intervals. Stimuli adapted from Nicol and Swinney (1989).

5.5.2 Results

Results by model for English are given in Figure 5.2. Statistical analyses were conducted via linear-mixed effects models.[11] Additionally pairwise t-tests were conducted to assess differences between the experimental conditions.[12]As with the prior experiment, we interpreted our results relative to which stimuli groups differ statistically. That is, do stimuli with gender agreeing syntactically licit noun phrases pattern together?

[11]We fit models to predict the surprisal of the pronoun *him* with interactions between the genders of all three antecedent nouns (i.e. NP1 *told* NP2 *that* NP3 *would blame him for the recent injury*). with random slopes by item for the gender of each noun antecedent. For the LSTM models, random slopes by model for the gender of each noun antecedent were also included. Random effects were removed if the statistical model failed to converge.

[12]To correct for multiple comparisons an adjusted p-value of 0.0003 was used.

The LSTMs and TransformerXL showed no interactions between antecedents instead masculine gender for any of the noun phrases (or just the grammatically licit ones for TransformerXL) lowered the surprisal of the pronoun him. The remaining models showed a more complex pattern with interactions between all three noun phrases (suggesting grammatically illict noun phrases influence the surprisal of *him*, contrary to humans). For GPT-2 XL, BERT, and RoBERTa, the experimental conditions MMM, MMF, MFF, and FMF all patterned together and were the least surprising. For GPT-2 XL the other experimental conditions grouped together, resulting in a surprisal ranking of {MMM, MMF, MFF, FMF} < {MFM, FMM, FFF, FFM}. Ignoring MMM, GPT-2 XL favored conditions where the final noun phrase did not agree with the pronoun and at least one of the grammatically licit nouns agreed in gender with the pronoun.

BERT had a surprisal ranking of {MMM, MMF, MFF, FMF} < {MFM, FMM} < {FFM, FFF}. That is, similarly to GPT-2 XL, BERT favored agreeing with at least one grammatically licit noun and disagreeing with the lower noun, then favored agreeing with one grammatically licit noun and agreeing with the lower noun, and finally agreeing with no grammatically licit nouns. RoBERTa had a surprisal ranking of {MMM, MMF, MFF, FMF} < {MFM, FMM, FFF} < FFM. That is, RoBERTa was similar to BERT except for preferring to agree with no nouns (FFF) over agreeing with only the lowest noun (FFM).

5.5.3 Discussion

Neural model behavior was more varied in the case of three antecedents than in the case of two antecedents. The LSTMs and TransformerXL showed a general effect of masculine agreement. For LSTMs this amounted to any masculine antecedent in

any position lowered the surprisal of him, while for TransformerXL only masculine gender on grammatically licit antecedents lowered surprisal. That is there was only very limited evidence for knowledge of Principle B for TransformerXL and none for the LSTMs.

GPT-2 XL, BERT, and RoBERTa showed, in some contexts, an ungrammatical match effect (a similar effect was discussed above, see Section 5.4.3). Recall, that in the prior section it was unclear whether neural models preferred to agree with grammatical licit antecedents or just the subject (or first noun). The results from the three antecedent experiment, suggests that, for GPT-2 XL, BERT, and RoBERTa, coreference preferences are broader than just a general preference for the first noun.

As with two antecedents, we found that the penalty for agreement with ungrammatical antecedents was not incurred when the grammatical antecedents agreed with the pronoun (i.e. MMM did not pattern with other conditions where the more local noun agreed in gender with the pronoun). Across GPT-2 XL, BERT, and RoBERTa, we saw behavior akin to a conditional constraint. Contexts where the ungrammatical antecedent agreed with the pronoun were dispreferred. This behavior has some correspondence to Principle B.

While suggesting a more complex pattern than proposed after the preceding section, we can nonetheless account for model behavior with relatively simple constraints. Namely, (i) agree in gender with a preceding noun phrase, and (ii) do not agree with the most recent noun. This has overlapping distribution with Principle B with an additional penalty for having no antecedents (i.e. penalizing FFF, despite no violation of Principle B). We turn to this point again in the General Discussion (see Section 5.8).

As with the other chapters in this dissertation, we turn to the interaction of linguistic processes as a way of validating whether the model behavior is truly tracking the grammatical generalization. In particular, we investigated the relationship between Principle B and cataphoric pronouns (that is, pronouns which agree with upcoming linguistic material; e.g., *While he was eating, Fig looked around*) for which psycholinguistic evidence suggests Principle B is immediately used to constrain human predictions (Kush and Dillon, 2021). To look ahead, we find that models do not use Principle B for forward prediction, suggesting Principle B is not fully learned by neural models.

5.6 Predictive Processing with Cataphora

The preceding experiments focused on backward anaphora, that is the gender of the pronoun is constrained by the gender of preceding nouns. While we found some preliminary evidence that neural models could behave in accordance with Principle B (though the immediate constraints on processing demonstrated in humans were not attested with neural models), it remains unclear whether the models have actually abstracted something like Principle B.

In order to explore this more fully, we looked to the interaction between Principle B and the prediction of upcoming material. It seems reasonable, given that neural models are explicitly trained to predict upcoming linguistic material, to compare the predictions of humans and neural models. Namely, we investigated whether cataphoric pronouns constrained predictions in accordance with Principle B for neural models, as they do for humans (Kush and Dillon, 2021).

As the interaction involves at least two processes, (i) predicting nouns based on

the gender of the cataphoric pronoun, and (ii) knowledge of Principle B, we began by validating whether neural models constrain their predictions due to cataphoric pronouns at all. We utilized the stimuli from van Gompel and Liversedge (2003), which demonstrated that cataphoric pronouns constrained the prediction of subject nouns for humans. That is, a masculine cataphoric pronoun lead to a mismatch effect when the subject was feminine (and similarly for feminine pronouns and masculine subjects, in addition to number mismatches). We found evidence that a subset of neural models did pattern like humans in predicting subjects with the gender of the cataphoric pronoun.

5.6.1 Stimuli

The stimuli are drawn from Experiment 1 in van Gompel and Liversedge (2003).[13] They explored whether cataphoric pronouns (i.e. pronouns appearing before their antecedents) constrain the prediction of upcoming nominals. Consider the following:

(10) a. When he was off work, the barman pestered the waitress all the time.
b. When he was off work, the waitress pestered the barman all the time.
c. When she was off work, the barman pestered the barman all the time.
d. When she was off work, the waitress pestered the barman all the time.

In (10-a) and (10-d) the cataphoric pronoun agrees in gender with subject, while it mismatches in (10-b) and (10-c). van Gompel and Liversedge (2003) found that human readers had a preference for linking the pronoun and the subject, and thus, reading times were slowed when there was a gender mismatch. A

[13]The stimuli templates are given in Appendix C.3.

similar result was obtained with number (e.g., *When he/they was/were off work, the barman/barmen...*). In the present study, 32 such stimuli were tested, each focusing on a gender contrast. Minor modifications were made when nouns were outside the vocabulary of the neural models.

5.6.2 Results

Results by model for English are given in Figure 5.3. Statistical analyses were conducted via linear-mixed effects models.[14] BERT, RoBERTa, and TransformerXL all showed increased surprisal when the subject did not agree with cataphoric pronoun (e.g., *When she felt sad, the man...* was more surprising than *When she felt sad, the woman...*). RoBERTa showed an additional effect of the pronoun gender, with generally greater surprisal when the pronoun was masculine. The LSTMs only had a mismatch effect for cataphora and subjects when the pronoun was masculine. GPT-2 XL showed no significant effects, suggesting that it did not predict the gender of the subject based on the gender of the cataphoric pronoun.

5.6.3 Discussion

We found evidence that BERT, RoBERTa, and TransformerXL all conditioned their predictions of subjects on the gender of preceding cataphoric pronouns (i.e., given *While he was swimming, the*, *man* was less surprising than *woman*). The LSTMs showed constrained predictions for only masculine pronouns, while GPT-2 XL showed no effects at all. We might then expect that BERT and RoBERTa (which

[14]We fit models to predict the surprisal of the subject with interactions between agreement between the cataphoric pronoun with the subject (Match vs. Mismatch) and the gender of the pronoun (*he* vs. *she*) with random intercept for items (and models in the case of the LSTMs).

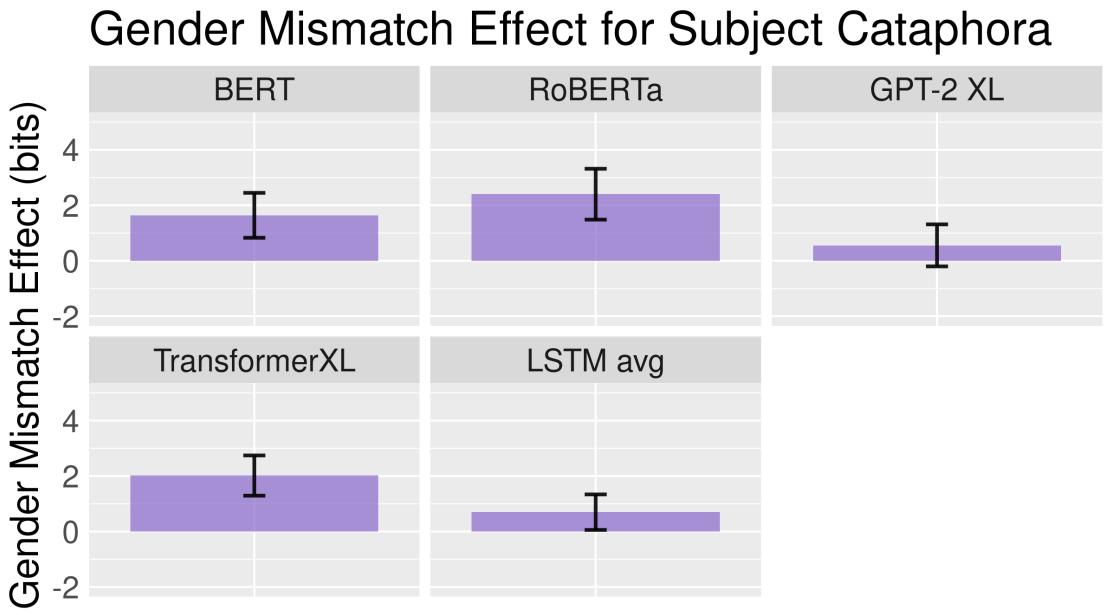

Figure 5.3: GMME by model for subject following a cataphoric subject pronoun (e.g., *While he was working, the (man/woman)...*). Error bars are 95% confidence intervals. Stimuli adapted from van Gompel and Liversedge (2003).

showed some behavior in line with Principle B in prior experiments) should show an interaction between subject prediction and Principle B. We turn to investigations of this below.

5.7 Interaction between Principle B and Predictive Processing

Recall, that we are interested in whether Principle B was learned by neural models. We found preliminary evidence in the context of pronouns conditioned on preceding context (though, it is worth repeating that the behavior of neural models diverged

from what we see in humans immediate processing). As with the prior chapters, we turned to the interaction of linguistic processes as a means of investigating the degree to which linguistic behavior in neural models is human-like. Above, we found evidence that models can condition the prediction of subject nouns on the gender of cataphoric pronouns (i.e. preferring nouns like *man* after *while he drove, the...*).

Kush and Dillon (2021) demonstrated that, in addition to predicting the gender of the subject based on the gender of the cataphoric pronouns, predictions were modulated by whether Principle B allowed coreference between the cataphoric pronoun and the subject (e.g., *him* cannot corefer with *Bob* in *While driving him, Bob ate a sandwich*). We replicated the experiments from that paper with neural models, finding no evidence that neural models condition their future predictions on Principle B.

5.7.1 Stimuli

The stimuli were drawn from Kush and Dillon (2021), which explored whether Principle B constrained cataphoric pronouns.[15] Consider:

(11) a. While reading **him** a bedtime story, William gently gestured to Luis to turn off the lights.

b. While reading **his grand kids** a bedtime story, William gently gestured to Luis to turn off the lights.

[15]The templates for the stimuli are given in Appendix C.4.

In (11), the critical contrast is bolded. Namely, in (11-a), the pronoun *him* cannot corefer with *William* because of Principle B.[16] While in (11-b), coreference between *his* and *William* is possible, and seems to be the more intuitive reading.

In addition to the above contrast, another version of (11-b) was used in a follow up experiment:

(12) While someone read **him** a bedtime story, William gently gestured to Luis to turn off the lights.

In (12), the pronoun **him** is also used. However, coreference with *William* is possible. Across both experiments, Kush and Dillon (2021) found a reading time slowdown when the cataphora and the subject disagreed in gender (i.e. a gender mismatch effect), but only when Principle B was not implicated (e.g., there was no gender mismatch effect in (11-a)). In what follows we investigated the extent to which neural models catch this empirical generalization.

In total, 24 such stimuli sets were investigated. Both masculine and feminine pronouns were investigated, in accordance with Kush and Dillon (2021), and right contexts were included for non-autoregressive models (i.e. the material after *William* in (11-a)). Proper names were changed to *the man* or *the woman*, since at least one neural model lacked each of the proper names.

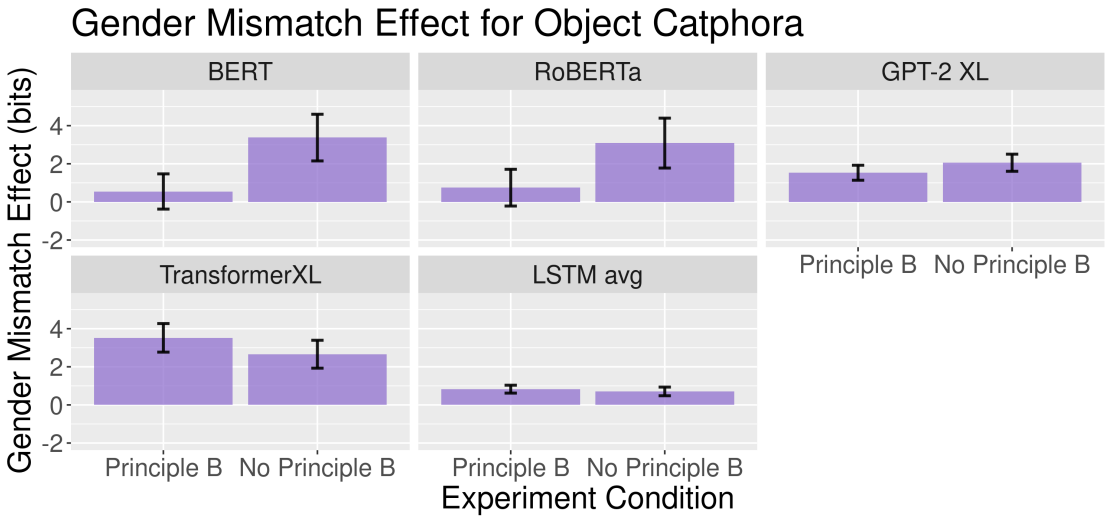

Figure 5.4: GMME for subject following a cataphoric object pronoun (e.g., *him*) for each neural model by whether Principle B applies (e.g., *Before offering (her/him) a fancy pastry, the man...* vs. *Before offering (his/her) son a fancy pastry, the man...*). Error bars are 95% confidence intervals. Stimuli adapted from Experiment 1 in Kush and Dillon (2021).

5.7.2 Results

Results by model for English are given in Figures 5.4 and 5.5. Statistical analyses were conducted via linear-mixed effects models.[17] Both experiments in this section targeted whether a mismatch gender effect for cataphora (that is greater surprisal when the cataphora mismatched in gender with the following subject) was modulated by Principle B. For humans, Kush and Dillon (2021) demonstrated that a gender mismatch effect only occurs when the subject can be grammatically co-indexed with the cataphora (i.e. Principle B immediately blocked any prediction about the gender of the subject).

[16]Obligatory control of PRO in the adjunct is also implicated. We abstract from the relevant syntactic analysis here, and instead focus on the empirical findings. See Kush and Dillon, 2021.

[17]We fit models to predict the surprisal of the subject with interactions between agreement between the cataphoric pronoun with the subject (Match vs. Mismatch) and whether Principle B was active or not with random intercept for items (and models in the case of the LSTMs).

Starting with the first experiment (where the presence or absence of a Principle B violation is modulated by the form of the pronoun, object vs. possessive), BERT and RoBERTa patterned like humans, only demonstrating a mismatch effect when there was no Principle B constraint. GPT-2 XL, TransformerXL and, the LSTMs only demonstrated main effects of cataphora agreement, with greater surprisal when the subject mismatched with the cataphoric pronoun (cf. GPT-2 XL in this experiment where cataphora do predict the subject with the prior experiment where no such effect was found). That is, on first glance, it may seem like BERT and RoBERTa are human-like, while the other models are not. Recall, however, that the cataphora differ in form depending on the presence or absence of a Principle B violation. In other words, these two models may be tracking the form of the pronoun (similarly the above section used subject pronouns, so it could be that for these models, *his* and *he* as cataphora constrain future noun phrases while *him* does not).

That brings us to the second experiment in Kush and Dillon (2021) which used the same pronoun form (*him* and *her*) while modulating the presence of Principle B violations (see Figure 5.5). Here, RoBERTa and BERT showed no effect of Principle B, in fact showing no gender mismatch effect at all. GPT-2 XL, TransformerXL, and the LSTMs showed the same generalization as with the first experiment, they demonstrated a mismatch effect in both experimental conditions with no differences with the presence of a Principle B violation. We return to a more nuanced comparison to human reading times in the general discussion.

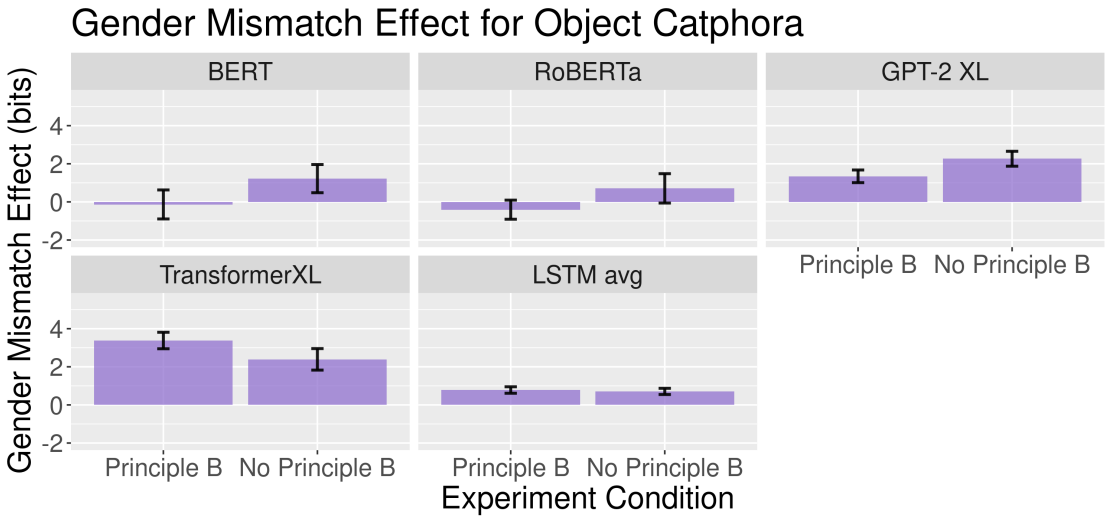

Figure 5.5: GMME for subject following a cataphoric object pronoun (e.g., *him*) for each neural model by whether Principle B applies (e.g., *Before offering her/him a fancy pastry, the man...* vs. *Before anyone offered her/him a fancy pastry, the man...*). intervals. Error bars are 95% confidence intervals. Stimuli adapted from Experiment 2 in Kush and Dillon (2021).

5.7.3 Discussion

We replicated the two experiments from Kush and Dillon (2021) using neural models. Despite apparent evidence for prediction modulated by Principle B for BERT and RoBERTa in the first experiment, the second experiment demonstrated that no neural model behaves in accordance with Principle B. That is, BERT and RoBERTa did not predict the gender of the subject with any object pronouns even when coreference was grammatical (in contrast to the results for subject pronouns in Section 5.6). For GPT-2 XL, TransformerXL, and the LSTMs, the subject was more surprising when the cataphoric pronoun differed in gender, regardless of whether Principle B blocked agreement. These results are in contrast to humans, who only show a gender mismatch effect when coreference between the subject and

the pronoun was not blocked by Principle B.

We return to a broader discussion of these results, and the results of the experiments throughout the chapter below. However, it appears that neural models learn to behave in accordance with Principle B only in specific contexts. That is, there doesn't appear to be any model which behaves in line with humans across all these experiments.

5.8 General Discussion

As with the preceding chapters, we have challenged the view that neural models of language learn human-like syntactic knowledge. While some models capture aspects of Principle B in canonical structures (i.e. backward coreference relations), we found that in more complex environments Principle B did not constrain neural model predictions. That is, Principle B in humans is a generalized (and abstract) constraint on coindexation that can be implicated by a number of different surface configurations, while neural models appear to learn more specific constraints on pronominal agreement. The narrowness of supposed grammatical knowledge in neural models has been a constant theme in this dissertation, and we return to this point in the conclusion. For the moment, however, we can look more closely at the inability of neural models to capture pronominal agreement effects in a human-like manner.

We have been implicitly comparing the behavior of neural models to the widely observed gender mismatch effect (GMME) in humans. For humans, this amounts to a reading time slow down (or cost in processing) which "indexes the comprehender's surprise at finding a non-antecedent where they expected to encounter the

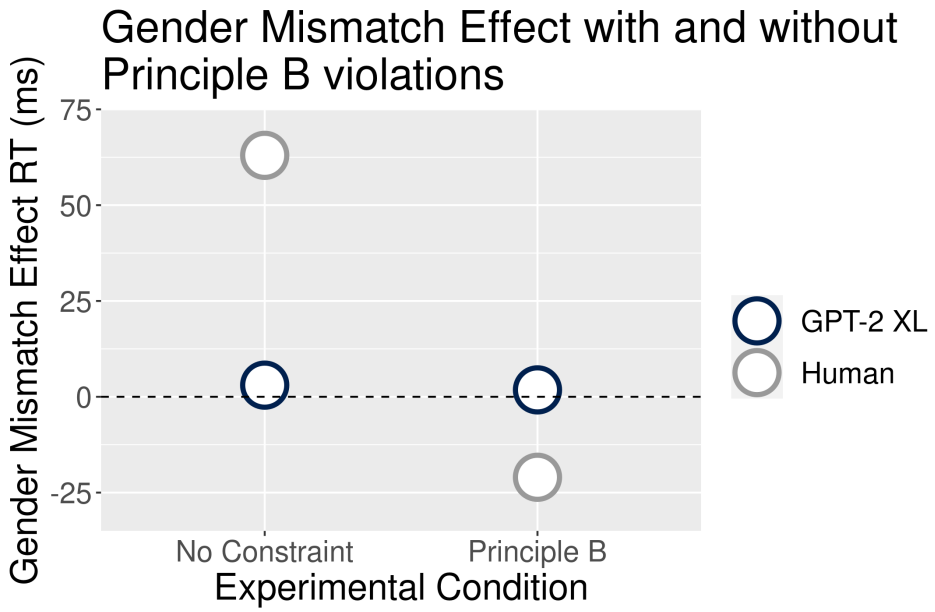

Figure 5.6: Mean gender mismatch effect for humans and GPT-2 XL for Experiment 2 in Kush and Dillon (2021). That is, the difference between mismatching cataphora by condition (e.g., *Before offering her son a pastry, the man...* vs. *Before offering his son a pastry, the man...*). GPT-2 XL predicted difference in reading times were obtained by fitting a model predicting self-paced reading times in the Natural Stories Corpus (Futrell et al., 2018a) with GPT-2 XL surprisal (following the method in van Schijndel and Linzen (2018a)).

antecedent" (Kush and Dillon, 2021, p.2). For Experiment 2 in Kush and Dillon (2021), this amounted to a difference in reading time of 63 ms for mismatches between the pronoun and the subject when Principle B was not active. In fact, when Principle B blocked coreference, there was a reverse gender mismatch effect of -21 ms (i.e. gender agreement between the pronoun and the subject had greater reading times). Given that surprisal has been shown to be linearly correlated with human reading times, we can ask whether the surprisal assigned by a neural model captured the effect size (Hale, 2001; Levy, 2008; Smith and Levy, 2013).

Following van Schijndel and Linzen (2018a), we fit a linear mixed effects model

relating the by-word surprisal of GPT-2 XL to the by word self-paced reading times reported in Futrell et al. (2018a).[18] The resultant statistical model found that every bit of surprisal correlated with an increase of 1.34 ms in reading times. Applying this to the predicted difference in surprisal for the conditions in Experiment 2 (detailed above) gives the predicted GMME for GPT-2 XL. The GMME for humans and GPT-2 XL is given in Figure 5.6.

As is visually apparent, even in cases where GPT-2 XL qualitatively matches the behavior in human processing (i.e. when Principle B does not block coreference), GPT-2 XL is far from capturing the estimated effect size in human processing. Similar GMMEs are reported for backward anaphora. For example, Chow et al. (2014) appears to have an effect size of around 60-70ms.[19] However, GPT-2 XL has a similar GMME of around 4 ms (when holding constant the mismatch in gender with the embedded subject).

Neural models seem to under-predict the processing cost of GMME. Similar results have been obtained in prior work for non-pronominal constructions, suggesting a broader inability for surprisal measures from neural models to capture the processing cost of grammatical violations (van Schijndel and Linzen, 2018a, 2021; Wilcox et al., 2021b; Paape and Vasishth, 2022).

We might follow the proposal in Wilcox et al. (2021b) and apply a scalar term to the predicted processing cost from surprisal. They find that this scalar term results in a significant improvement in fit to human behavioral measures. Crucially, however, this approach cannot handle the present findings. The predicted GMME for GPT-2 XL is in the wrong direction for the interaction with Principle B (humans

[18]We also included a main effect of word length and fit a maximal model with with by subject and by item random slopes.

[19]This is from Table 2. The estimated GMME sizes is not reported in the paper and the full results are not available online.

had a negative GMME for such stimuli). That is, we may alleviate the difference in effect size between neural models and humans in grammatical constructions, but this leaves unaddressed the problem of Principle B.

Turning again to the status of human-like Principle B in neural models, our results suggest that a mismatch remains. That is, neural models fail to learn both the immediate processing behavior of humans (i.e. neural models do not immediately restrict antecedents to those that are grammatical), and the relation between Principle B and other surface configurations (i.e. the interaction of Principle B and cataphora). We suggested that two constraints may account for the behavior of the "successful" neural models, 1) pronouns must agree in gender with some preceding noun, and 2) object pronouns cannot agree with the most recent noun.

It is relatively easy to imagine that the first constraint is just capturing the fact that, in linguistic data, pronouns rarely occur out of the blue. The second constraint, which appears to gesture at something like Principle B, may also be epiphenomenal. It could well be that the sequence masculine subject followed by a verb and a masculine object pronoun is less common than with feminine subjects. This could follow from a general avoidance in written text of any possibly ambiguous strings (here an ungrammatical reading is being avoided). We leave to further work investigating whether this holds, but nonetheless it suggests that surface forms may encode grammatical distinctions, which allows models to capture some relevant grammatical restrictions in limited contexts, while neural models fail to learn the underlying grammatical knowledge that drives such surface distinctions. Such a proposal was elaborated, to some degree, in both of the preceding chapters. We return to this point in the conclusion below.

Ultimately, this chapter has shown that neural models fail to capture both

human-like online coreference processing, and also fail to capture the full suite of behaviors associated with Principle B. As discussed above, even when models qualitatively pattern like humans, they miss the estimated effect size by an order of magnitude. That is, while neural models may appear to capture limited aspects of human linguistic processing, there is still work to be done on resolving mismatches between the two.

CHAPTER 6

CONCLUSION

This dissertation focused on the relationship between linguistic behavior in neural models and humans, and more broadly what linguistic behaviors follow from unstructured linguistic data. While concurrent work has suggested that neural models acquire aspects of human-like linguistic knowledge, Chapters 3–5 have shown that such knowledge is limited and ultimately deeply different from the behaviors of humans.

6.1 Summary

Through three case studies i) implicit causality, ii) ambiguous relative clause attachment, and iii) the interaction between Principle B and coreference processing, this work has provided evidence for three mismatches (not necessarily mutually exclusive):

1. Constraint ranking (and abstraction)
2. Production and comprehension
3. Parsing mechanisms

Each mismatches links a difference between neural models and humans to a fundamental limitation of linguistic data. The first mismatch notes that human-like linguistic systems contain intricate relationships between individual linguistic processes. Critically, data only provide the surface manifestations of these processes which can obscure their underlying relations, leading models astray. The second

mismatch captures the fact that training data are production data not comprehension data. Therefore, production biases will dominate model behavior. The third mismatch indicates that the human parser employs strong constraints that are not necessarily evident in data. Thus, models can, and do, develop incremental processing behaviors that diverge from humans. In what follows, the evidence for each mismatch is summarized and the contents of the mismatch detailed.

6.1.1 Constraint ranking

In investigating implicit causality (Chapter 3), we found evidence that neural models of languages other than English fail to learn IC despite cross-linguistic similarity in humans, and in fact, even neural models of English fail to learn the full suite of behaviors associated with implicit causality. Rather than failing on a given language at random, we suggested that the failure of neural models was principled: cross-linguistically, models fail depending on which other linguistic processes are active.

Specifically, in the case of IC, we showed that the failure of a neural model to behave in accordance with an IC bias occurred when pro-drop (a process which allows empty subject positions) was active in a language. Pro-drop disprefers overt pronouns, while IC prefers coreference with an antecedent (as evidenced by the presence of a pronoun agreeing in gender with an antecedent in our experiments). Thus these processes were in direct competition (i.e. they both affect pronouns).

Alleviating this competition by removing pro-drop via fine-tuning (for Spanish and Italian), or instantiating this competition by adding a pro-drop constraint via fine-tuning (for English and Chinese) confirmed our conclusion: pro-drop obscures

underlying IC knowledge. We further suggest that this mismatch between neural models and humans points to a broader mismatch between constraint rankings in humans and neural models.

Neural models learn to rank constraints in accordance to their frequency in data, while humans can rank constraints relative to other considerations (such as faithfulness to some underlying representation or universal markedness). The overlapping distribution of an IC constraint and a pro-drop constraint, ultimately leads the model to favor the behavior which conforms to the most frequent of these constraints (see Section 3.5 for further discussion). In contrasts, these additional considerations beyond the frequency of processes in data account for the robust IC effects for humans.

6.1.2 Production and comprehension

In Chapter 4, we investigated whether neural models of English and Spanish obtained the attachment preferences of speakers of these languages. While humans have differing attachment preferences, low attachment in English and high attachment in Spanish, neural models prefer low attachment in both languages. Careful investigation of construction specific attachment preferences also suggested that neural models fail to acquire the full range of attachment preferences demonstrated in English. That is, while a low attachment preference is globally preferred, there are cases of high attachment preference for English speakers. Neural models, on the other hand, seem to have an overwhelming preference for low attachment.

We hypothesized that neural models were tracking early parsing behaviors of humans, rather than their ultimate interpretation preferences. Studies with

Spanish speakers have suggested that number disambiguation favors low attachment in early processing (see Fernández, 2003). Additional experiments with Spanish demonstrated, as in the case of humans, that gender disambiguation yielded high attachment preferences in neural models. While further work must be done to explain these differences, our results point to an interesting relationship between the behavior of neural models and levels of linguistic processing in humans.

In sum, we found evidence that attachment preferences in neural models tended towards the dominate attachment preference in production data (a preference for low attachment). In other words, while Spanish speakers have a general interpretation preference for high attachment, and even English speakers have construction specific high attachment preferences, neural models learn to pattern with the more frequent attachment pattern in data (cf. Desmet et al., 2006). The training signal for neural models, then, provides evidence for production preferences (and maybe initial parsing behaviors, though see below), rather than evidence for interpretation preferences. This mirrors recent discussion in natural language processing concerning the limitation of text-based models and suggests a broader limitation of naive language models (see Bender and Koller, 2020; Bisk et al., 2020).

6.1.3 Parsing mechanisms

Building on the findings above, the final experiments of this dissertation (Chapter 5) explored the relationship between binding conditions (Chomsky, 1981) and immediate coreference behaviors. In establishing coreference, human grammar has a number of syntactic constraints, namely the Binding Principles. In Chapter 5, we explored Principle B, which has been shown in human experiments to immediately restrict parsing behaviors (e.g., Nicol, 1988; Chow et al., 2014; Kush and Dillon,

2021).

We found that, while neural models may behave in accordance with Principle B in some cases, neural models fail to capture both the general parsing behaviors of humans (i.e. they consider antecedents that humans do not) and also fail to capture the full suite of behaviors associated with Principle B (i.e. they learn a subset of what we call Principle B). It seems instead that neural models learn simple heuristics that mimic Condition B in specific contexts. This is in contrast to learning a more general constraint, corresponding to Principle B itself, which produces a range of surface behaviors. Future work should establish whether these heuristics follow directly from the distribution of pronouns in English training data, as well as broaden the investigation to other languages.

Ultimately, we claim that, at least some, aspects of parsing behavior follow not from training data, but from the architecture of the human parser (or human grammar). That is, while neural models have been related to existing theories of the human parser (Ryu and Lewis, 2021), there remain fundamental mismatches between humans and neural models. Similar thoughts are echoed by Fodor, who states that while one can build a parser, "what does *not* follow is that there is some way of constructing such systems [parsers] from the information given *in experience*" (Fodor, 1983, p. 35).

6.2 Superficialism and the Illusion of Grammatical Competence

In Chapter 2, we discussed the notion of superficialism, or the belief that all meaningful psychological distinctions can be made on the basis of behavior (see Rey, 2020). Within this dissertation, we assumed a form of superficialism (which is implicit in much related work), in which we assume all meaningful linguistic structures (or representations) can be learned on the basis of linguistic behavior. In other words, all meaningful linguistic structures are evidenced by surface contrasts (no process is fully opaque). This dissertation has revealed limitations of this world view via three case studies demonstrating neural models remain far from obtaining human-like linguistic behavior. Superficialism, however, remains a way of understanding the cases where models do learn more limited aspects of human behavior. Successful overlap between neural models and humans occurs in just those cases where surface contrasts alone (with no reference to meaning[1]) consistently distinguish the process.

While more work must be done to substantiate these claims (e.g., what makes a process consistent; what counts as a surface contrast), consider the case of implicit causality (Chapter 3). We have argued that competition between linguistic processes can obscure underlying linguistic knowledge in neural models (and presumably in people as is argued in generative linguistics). That is, models are constrained by the frequency of surface forms in language. In learning a linguistic process, then, the strength of the association between the linguistic process and the example sentences in a corpus modulates the ability of a neural model to learn a linguistic

[1]No reference to meaning is not meant to be a necessary condition. Instead, it just picks out the fact that neural models of language have no direct access to interpretations, so at present there seems to be no way such models could leverage meaning coupled with surface contrasts

abstraction in line with humans.

If an abstract representation consistently realizes a surface contrast, and this surface contrast reliably correlates with the abstract representation (to the exclusion of other representations), then a neural model may appear to learn a human-like behavior. Subject-verb agreement and implicit causality appear to have this desired correspondence at least in English, such that there is considerable overlap between the presence of certain subjects (or certain verbs) and the number on certain verbs (or the agreement on pronouns). This may lead us to propose that models learn something analogous to the human representation, which we think causes the surface behavior. However, auxiliary linguistic processes in different languages, in the case of implicit causality, can prevent neural models from learning the same structure (despite the ability of humans to do so in both languages). Thus, human linguistic abstractions are, seemingly, more robust than surface patterns may suggest.

In the field, we may be studying those linguistic processes which are most strongly realized in the surface forms of language, and thus overemphasizing the ability of neural models to learn human-like linguistic structure. Exploration of languages other than English and a broadening of the phenomena we investigate will, this dissertation suggests, continue to demonstrate how linguistically naive neural models fall short of human linguistic knowledge.

6.3 Linguistic Theory and Neural Models

> [Phonological] representations are not derived from the speech sounds by analytic procedures of segmentation, classification, extraction of physical features, and so forth, but are established and justified as part

> of the best theory for accounting ultimately for the general relation between sound and meaning of the I-language. (Chomsky, 1986, p.43)

This dissertation has discussed linguistic processes and linguistic representations in the evaluation of neural models of language, as do many papers in the growing literature interpreting neural models. It is critical to elaborate on the relationship between linguistic theory (and the processes and representations utilized in that domain) and the language used to describe the behavior of neural models, if we wish to understand both how neural models work, but also what neural models tell us about human linguistic knowledge. A few existing papers have attempted to do just this (e.g., Bommasani et al., 2021; Linzen and Baroni, 2021; Wilcox et al., 2021a). A common argument in such work is to suggest that behavioral overlap between neural models and humans is an argument against poverty of the stimulus claims in theoretical linguistics (we return to poverty of the stimulus arguments in the following section). This argument against poverty of the stimulus rests on the assertion that models learn human-like linguistic representations.

However, claims that models do learn human-like linguistic representations face, at least, two immediate challenges. The first points to a lack of systematicity: in a single language, a neural model may overlap with human behaviors for some phenomenon, but fail to obtain other behaviors which utilize this same phenomenon (see Sections 4.6 and 5.7). The second points to a strong reliance on surface patterns: across languages with the same phenomenon, a neural model may behave in accordance with humans in only a subset of these languages (see Section 3.4).

In linguistic theory, phenomena are both systematic (being utilized in the explanation of a number of contrasts) and cross-linguistically stable (i.e. accounts of multiple languages use the same underlying theoretical constructs). If models

are learning the same structures and representations detailed in linguistic theory, then presumably they should bear these same properties. This dissertation suggests that neural models do not do so, and therefore they are not learning the objects in linguistic theory.

This does not doom the study of neural models. As the quote at the beginning of this section suggests, representations in linguistic theory play an explanatory role within a theory (and are evaluated with criteria distinct from accounting for all the sentences in a given corpus). It is entirely possible (and quite likely) that surface driven models arrive at abstractions that are quite alien to those in linguistic theory.[2] We may begin to use neural models to inform our models of language (e.g., as "theories" of acceptability judgments; Baroni, 2022), if we explicitly interrogate these structures, and at least to some degree, schematize them. Otherwise, we gain no insight about language, or humans, from the investigation of model behaviors, and overlap between neural models and humans remains accidental.

We know that language is a hugely interactive system, with individual utterances influenced by many factors outside of a narrow conception of language (e.g., world knowledge, socio-indexical information, social conventions). Neural models are trained on the output of this multi-faceted system. It may, therefore, be that neural models arrive at the wrong mix of factors (perhaps learning more about the conventions of wikipedia articles and reddit forums, than the underlying generative system posited in theory). Chomskyan linguistic theory has attempted to isolate a core set of properties for formal study, leaving aside the broader creative use of language, which may be beyond our ability to study (see for example Chomsky, 1965, 1995, 2000). Combining insights from theory about the basic units of the

[2]In fact, Rey (2020) argues that linguistic representations are "intentional inexistents", entities which are "the 'things' mental states are 'of' or 'about', but do not actually exist." (Rey, 2020, p. 8).

language faculty with neural models trained on large amounts of examples may yield interesting insights into the difficult-to-study aspects of language like the creative use of language.

6.4 Neural Models and Poverty of the Stimulus: The View from Below

The poverty of the stimulus argument advanced famously by Chomsky and articulated by generative grammar points to the mismatch between the representations in our explanatory theories and the contents of typical linguistic experience. That is, typical linguistic experience fails to provide determinate evidence of the elements we believe are necessary to explain the knowledge native speakers possess (or in other words, experience is consistent with many possible hypotheses, not necessarily only the ones we consider within a single theory). This issue of indeterminacy is common to any scientific exploration of a topic which aims to abstract from the complexity of everyday experience more fundamental principles.

Nonetheless, the claim that linguistic knowledge goes beyond typical linguistic experience is still controversial within natural language processing. Many people continue to advance the position that experience is largely determinate of linguistic knowledge. This is, I would say, best articulated in the current intersection of linguistics (and psycholinguistics) with natural language processing (as mentioned throughout this thesis). The project, as perceived by many practitioners, is to reduce human competences (or, more neutrally, abilities) to some general statistical mechanism applied to the contents of experience. Human language, then, is no more than extremely accurate prediction tuned by large amount of experience.

There remains a critical error in reasoning from the behavior of neural models to human linguistic structures or representations. Similar behavior does not mean similar representations, so overlapping behavior is not necessarily pointing to anything that deeply shared between computational models and human knowledge (see Guest and Martin, 2021 for an extremely lucid explication of this point). Of course, there are other limitations. The reductionist approach fails to provide reasons for why languages have the curious properties that they do. Explanations of such properties, both within single languages and in cross-linguistic comparison, is a major goal of theoretical linguistics, so any accounting of language that fails to address considerations of this kind face an explanatory gap. In fact, the reductionist approach shifts the focus from the mind of a speaker, and instead seems to claim that language data itself provides all the evidence for these properties. Consider a world where IC was reliably and consistently surface evident in all of the worlds languages (in other words, neural models could learn IC equally well in all languages; cf. Chapter 3). The stability of the linguistic signal would not explain why IC verbs had the biases they had. It would merely provide evidence that they do.

This thesis presents yet another issue for reductionist approaches. One that is more empirically grounded. Namely, the 'linguistic' systems of computational models trained on unstructured experience face a number of seemingly fatal problems: i) the system is surface bound, and thus fails to acquire opaque processes, or processes with overlapping conditioning environments, in the same way as humans, ii) the system has no direct access to interpretation (or comprehension), and thus can only capture aspects of the human linguistic production system, and iii) the system lacks architectural constraints of human minds (and linguistic data do not seemingly provide enough evidence to acquire these constraints). These problems follow from careful comparison between computational models trained on other

languages and from careful investigation of a fuller set of behaviors associated with the same linguistic representation.

Returning to the topic of poverty of the stimulus, this dissertation suggests an approach to the disconnect between experience and knowledge, not from an articulation of a theory and its lack of correspondence in experience, but from an investigation of the types of systems that experience lends itself to. Put another way, this work complements what we might call a top down approach that articulates an explanatory theory (and demonstrates the lack of empirical evidence for the mechanisms and representations implicated in the theory), by taking seriously a bottom up approach that centers experience, and by showing how such an approach is critically limited. This, in turn, helps clarify the gap between the human capacity for language and the data we are exposed to.

The approach advocated here bears resemblance to points recently repeated by Chomsky in discussion of the Minimalist Program (see, for example, his UCLA Lectures). Namely, human linguistic generalizations are constrained by what our minds cognize. For example, human linguistic systems do not make reference to seemingly simpler generalizations like linear order, because linear order is not accessible to our faculty of language. This dissertation argues for an approach to computational linguistics that contributes to theoretical linguistics by demonstrating additional generalizations that could emerge from the data (and are in fact more easily evidenced by experience) but that, crucially, humans do not make.

6.5 Future Directions

This dissertation has argued that neural models of language mismatch with human linguistic knowledge in a number of ways. These mismatches appear to be more fundamentally related to properties of linguistic data, and thus point to larger mismatches between experience with language and knowledge of language. The studies reported here make a number of predictions worth investigating:

- Linguistic processes with overlapping distributions will be difficult for models to learn in a human-like fashion.
- High attachment preferences will be learned by neural models only in cases where the production data tracks with comprehension preferences. This could be fruitfully investigated by examining neural models of other Romance languages like Italian and French.
- Neural models with innate parsing restrictions in line with architectural restrictions (e.g., binding principles) will facilitate greater overlap between neural models and human behavior.

Ultimately, the present thesis suggests that detailing mismatches between humans and neural models tells us more about the capacity of computational models than what we can learn simply from finding overlap between the two. Moreover, this approach demonstrates concrete cases where the human mind goes beyond experience, and thus, suggests points of fruitful discussion between linguistic theory and neural models.

APPENDIX A

APPENDIX FOR IMPLICIT CAUSALITY

A.1 Verbs and Noun Pairs

The English IC verbs are given in Table A.4, Chinese in Table A.1 (with the verbs in pinyin as provided in Hartshorne et al. (2013)), Spanish in A.2, and Italian in A.3. Noun pairs used to generate stimuli are given in Table A.5.

Verb	IC	Verb	IC
xianmu	O	jidu	O
kelian	O	haipa	O
tongqing	O	taoyan	O
danxin	O	guanxin	O
ciji	O	anwei	O
anfu	O	jinu	O
guli	O	youhuo	S
chunu	S	jili	S
kunrao	S	guwu	S
xiyin	S	xinren	O
zunjing	O	peifu	O
zenghen	O	chongbai	O
xinteng	O	choushi	O
jingwei	O	huaiyi	O
zunzhong	O	anlian	O
xiangnian	O	daonian	O
ganji	O	huainian	O
baorong	O	xiuru	O
biandi	O	wuru	O
qifu	O	zanyang	O
zhichi	O	chaoxiao	O
weixie	O	saorao	O
gouyin	O	gufu	S
qipian	S	shuofu	S
qifa	S		

Table A.1: Chinese IC verbs and bias (S for subject-biased and O for object-biased) from Hartshorne et al. (2013).

Verb	IC	Verb	IC
alcanzó	S	contempló	O
demostró	S	descubrió	O
eligió	O	encontró	S
escuchó	O	levantó	O
llamó	S	miró	O
pagó	S	preguntó por	S
rompió	S	usó	S
visitó	O	aceptó	S
criticó	O	cuidó	O
debía	O	defendió	O
detuvo	O	escapó de	O
obedeció	O	protegió	O
rectificó	O	respondió	O
salvó	O	señaló	O
aburrió	S	afectó	S
alteró	S	amenazó	O
animó	O	asombró	S
calmó	O	conmovió	S
distrajo	S	fascinó	S
impresionó	S	invitó	S
recordó	O	satisfizo	S
admiró	O	confió en	O
consideró	O	despreció	O
se enamoró de	O	enjuició	O
envidió	O	estimó	O
imaginó	S	molestó	O
se olvidó de	S	perdonó	O
prefirió	O	preocupó	O
quiso	O	reconoció	O
respetó	O	temió	O
valoró	O		

Table A.2: Spanish IC verbs and bias (S for subject-biased and O for object-biased) from Goikoetxea et al. (2008).

A.2 Expanded Results (including mBERT)

The full details of the pairwise *t*-tests conducted for the present study are given below (including the results for mBERT). The results for English models are in

Verb	IC	Verb	IC
accompagna in macchina	O	da spintone a	S
pettina	O	abbraccia	O
schiaffeggia	S	balla con	O
bacia	O	trattiene	O
telefona a	O	fa solletico a	S
lusinga	O	disobbedisce a	S
aiuta	S	imbroglia	S
inganna	S	tradisce	S
sorprende	S	salva	S
incoraggia	O	fa male a	S
affescina	S	piace a	S
diverte	S	crea problemi a	S
delude	S	spaventa	S
stupisce	S	infastidisce	S
annoia	S	mette in guardia	O
odia	O	si fida di	O
compatisce	O	teme	O
nota	O	stima	O
rispetta	O	ricorda	O
detesta	O	invidia	O

Table A.3: Italian IC verbs and bias (S for subject-biased and O for object-biased) from Mannetti and De Grada (1991).

Table A.6, for Chinese models Table A.7, for Spanish models Table A.8, and Italian models Table A.9.

A.3 Additional Fine-tuning Training Information

The full breakdown of pronouns added or removed in the fine-tuning training data are detailed below. English can be found in Table A.10, Chinese can be found in Table A.11, Spanish can be found in Table A.12, and Italian can be found in Table A.13.

Verb	IC	Verb	IC	Verb	IC	Verb	IC	Verb	IC
abandoned	S	acclaimed	O	accompanied	O	accused	S	admired	O
admonished	O	adored	O	advised	O	affected	S	aggravated	S
agitated	S	alarmed	S	alienated	S	amazed	S	amused	S
angered	S	annoyed	S	answered	O	apologized to	S	appalled	S
applauded	O	appreciated	O	approached	S	astonished	S	astounded	S
attracted	S	avoided	S	baffled	S	banished	O	battled	S
believed	O	betrayed	S	bewildered	S	blamed	O	blessed	O
bored	S	bothered	S	called	S	calmed	O	calmed down	O
captivated	S	carried	O	castigated	O	caught	O	cautioned	O
celebrated	O	censured	O	charmed	S	chased	O	cheated	S
cheered	O	cherished	O	chilled	S	comforted	O	commended	O
compensated	S	complemented	O	complimented	O	concerned	S	condemned	O
confessed to	S	confided in	S	confounded	S	confused	S	congratulated	O
consulted	S	corrected	O	corrupted	S	counseled	O	courted	S
criticized	O	dated	S	debated with	S	deceived	S	decried	O
defied	S	delighted	S	denounced	O	deplored	O	deprecated	O
derided	O	deserted	S	despised	O	detested	O	disappointed	S
discouraged	S	disgruntled	S	disliked	O	disobeyed	S	disparaged	S
distracted	S	distressed	S	distrusted	O	divorced	O	dominated	S
dreaded	O	dreamed about	S	echoed	S	embraced	S	employed	O
encouraged	O	enlightened	O	enraged	S	enticed	S	escorted	O
esteemed	O	exalted	O	exasperated	S	excited	S	excused	O
exhausted	S	fascinated	S	favoured	O	feared	O	fed	O
filmed	O	flattered	S	floored	S	followed	S	fooled	S
forgave	S	forgot	O	fought	S	freed	O	frightened	S
frustrated	S	grabbed	O	grazed	S	greeted	O	guided	O
hailed	O	harassed	S	harmed	S	hated	O	haunted	S
helped	O	hired	O	hit	O	honoured	O	hugged	S
hurt	S	idolized	O	incensed	S	infuriated	S	inspired	S
instructed	O	insulted	S	interrupted	S	intimidated	S	intrigued	S
irritated	S	killed	S	kissed	S	lauded	O	laughed at	O
led	O	left	S	lied to	S	liked	O	loathed	O
loved	O	married	S	met	S	missed	O	mocked	O
mourned	O	moved	O	noticed	O	ordered around	S	pacified	O
pardoned	O	passed	O	penalized	O	persecuted	O	picked up	O
plagued	S	played	S	played with	O	pleased	S	praised	O
prized	O	prosecuted	O	protected	O	provoked	S	punished	O
pursued	S	questioned	S	reassured	O	rebuked	O	relaxed	S
relished	O	repaid	S	repelled	S	reprimanded	O	repulsed	S
resented	O	respected	O	revered	O	revitalized	S	revolted	S
rewarded	O	ridiculed	O	rushed to	O	saluted	O	scared	S
scolded	O	scorned	O	shadowed	S	shocked	S	shook	O
snubbed	S	staggered	S	stared at	O	startled	S	stimulated	S
struck	O	sued	O	supported	O	surprised	S	tailed	S
telephoned	S	thanked	O	tolerated	S	took away	O	tormented	S
tracked	S	trailed	S	treasured	O	troubled	S	trusted	O
unnerved	S	unsettled	S	uplifted	O	upset	S	valued	O
venerated	O	victimized	S	vilified	S	visited	O	wanted	O
warned	O	welcomed	O	worried	S	worried about	O	worshipped	O
wounded	S	yearned for	S	yelled at	O				

Table A.4: English IC verbs and bias (S for subject-biased and O for object-biased) from Ferstl et al. (2011).

Noun1	Noun2	Lang	Noun1	Noun2	Lang
man	woman	EN	男人 (man)	女人 (woman)	ZH
boy	girl	EN	男孩 (boy)	女孩 (girl)	ZH
father	mother	EN	父 (father)	母 (mother)	ZH
uncle	aunt	EN	叔叔 (uncle)	姨 (aunt)	ZH
husband	wife	EN	丈夫 (husband)	妻子 (wife)	ZH
actor	actress	EN	王子 (prince)	公主 (princess)	ZH
prince	princess	EN	王 (king)	女王 (queen)	ZH
waiter	waitress	EN	儿子 (son)	女儿 (daughter)	ZH
lord	lady	EN	哥哥 (elderbrother)	姐姐 (eldersister)	ZH
king	queen	EN	弟弟 (youngbrother)	妹妹 (youngsister)	ZH
son	daughter	EN	孫兒 (grandson)	孫女 (granddaughter)	ZH
nephew	niece	EN	姪兒 (nephew)	姪女 (niece)	ZH
brother	sister	EN	叔父 (uncle)	姑母 (aunt)	ZH
grandfather	grandmother	EN	堂弟 (cousin)	堂妹 (cousin)	ZH
hombre	mujer	ES	uomo (man)	donna (woman)	IT
chico	chica	ES	ragazzo (boy)	ragazza (girl)	IT
padre	madre	ES	padre (father)	madre (mother)	IT
tio	tia	ES	zio (uncle)	zia (aunt)	IT
esposo	esposa	ES	marito (husband)	moglie (wife)	IT
actor	actriz	ES	attore (actor)	attrice (actress)	IT
príncipe	princesa	ES	principe (prince)	principessa (princess)	IT
camarero	camarera	ES	cameriere (waiter)	cameriera (waitress)	IT
señor	dama	ES	signore (lord)	signora (lady)	IT
rey	reina	ES	re (king)	regina (queen)	IT
hijo	hija	ES	figlio (son)	figlia (daughter)	IT
sobrino	sobrina	ES	nipote (nephew)	nipote (niece)	IT
hermano	hermana	ES	fratello (brother)	sorella (sister)	IT
abuelo	abuela	ES	nonno (grandfather)	nonna (grandmother)	IT

Table A.5: Nouns used to create stimuli for English, Chinese, Spanish, and Italian. The Spanish and Italian nouns share the same translation.

model	O-O μ	O-S μ	CI	p	S-O μ	S-S μ	CI	p
BERT	0.72	0.52	[0.19,0.21]	$< 2.2e^{-16}$	0.13	0.26	[0.12,0.13]	$< 2.2e^{-16}$
BERT_BASE	0.75	0.52	[0.11,0.13]	$< 2.2e^{-16}$	0.06	0.15	[0.08,0.09]	$< 2.2e^{-16}$
BERT_PRO	0.51	0.52	[0.14,0.15]	$< 2.2e^{-16}$	0.04	0.11	[0.06,0.07]	$< 2.2e^{-16}$
RoBERTa	0.57	0.41	[0.15,0.17]	$< 2.2e^{-16}$	0.31	0.43	[0.11,0.13]	$< 2.2e^{-16}$
RoBERTa_BASE	0.58	0.45	[0.11,0.13]	$< 2.2e^{-16}$	0.31	0.37	[0.07,0.08]	$< 2.2e^{-16}$
RoBERTa_PRO	0.35	0.23	[0.11,0.13]	$< 2.2e^{-16}$	0.16	0.19	[0.03,0.04]	$< 2.2e^{-16}$
mBERT	0.58	0.59	[-0.003,-0.01]	0.001	0.29	0.28	[-0.002,-0.01]	0.0002

Table A.6: Results from pairwise *t*-tests for English across the investigated models. O-O refers to object antecedent after object-biased IC verb and O-S to object antecedent after subject-biased IC verb (similarly for subject antecedents S-O and S-S). CI is 95% confidence intervals (where positive is an IC effect). BERT_BASE and BERT_PRO refer to models fine-tuned on baseline data and data with a pro drop process respectively.

model	O-O μ	O-S μ	CI	p	S-O μ	S-S μ	CI	p
BERT	0.41	0.39	[0.003,0.05]	0.00003	0.11	0.22	[0.09,0.12]	$< 2.2e^{-16}$
BERT_BASE	0.53	0.47	[0.03,0.08]	$2.2e^{-6}$	0.12	0.25	[0.11,0.14]	$< 2.2e^{-16}$
BERT_PRO	0.23	0.23	[-0.02,0.02]	0.94	0.04	0.11	[0.05,0.07]	$< 2.2e^{-16}$
RoBERTa	0.40	0.33	[0.04,0.08]	$1.16e^{-9}$	0.06	0.12	[0.04,0.06]	$< 2.2e^{-16}$
RoBERTa_BASE	0.52	0.46	[0.04,0.08]	$8.4e^{-7}$	0.05	0.11	[0.05,0.07]	$< 2.2e^{-16}$
RoBERTa_PRO	0.32	0.29	[0.002,0.06]	$7e^{-6}$	0.03	0.06	[0.02,0.04]	$< 2.2e^{-16}$
mBERT	0.08	0.07	[0.01,0.03]	$2e^{-6}$	0.08	0.06	[-0.009,-0.002]	$1.3e^{-5}$

Table A.7: Results from pairwise *t*-tests for Chinese across the investigated models from Cui et al. (2020). O-O refers to object antecedent after object-biased IC verb and O-S to object antecedent after subject-biased IC verb (similarly for subject antecedents S-O and S-S). CI is 95% confidence intervals (where positive is an IC effect). BERT_BASE and BERT_PRO refer to models fine-tuned on baseline data and data with a pro drop process respectively.

model	O-O μ	O-S μ	CI	p	S-O μ	S-S μ	CI	p
BERT	0.53	0.46	[0.04,0.09]	$1.4e^{-8}$	0.05	0.05	[0.0007,0.01]	0.03
BERT_BASE	0.37	0.30	[0.05,0.08]	$8e^{-12}$	0.03	0.03	[-0.004,0.007]	0.61
BERT_PRO	0.73	0.67	[0.05,0.07]	$< 2.2e^{-16}$	0.16	0.13	[0.01,0.03]	$1.2e^{-7}$
RoBERTa	0.09	0.10	[-0.008,-0.01]	0.03	0.06	0.06	[0.0007,0.007]	0.02
RoBERTa_BASE	0.06	0.06	[-0.005,-0.002]	0.0002	0.04	0.04	[-0.0003,0.004]	0.09
RoBERTa_PRO	0.48	0.48	[-0.03,0.01]	0.42	0.29	0.30	[-0.006,0.02]	0.24
mBERT	0.12	0.11	[0.001,0.01]	0.02	0.02	0.02	[-0.0002,-0.002]	0.03

Table A.8: Results from pairwise t-tests for Spanish across the investigated models from Cañete et al. (2020) and Romero (2020). O-O refers to object antecedent after object-biased IC verb and O-S to object antecedent after subject-biased IC verb (similarly for subject antecedents S-O and S-S). CI is 95% confidence intervals (where positive is an IC effect). BERT_BASE and BERT_PRO refer to models fine-tuned on baseline data and data with a pro drop process respectively.

model	O-O μ	O-S μ	CI	p	S-O μ	S-S μ	CI	p
BERT	0.21	0.19	[0.005,0.03]	0.004	0.09	0.11	[0.01,0.03]	$1.3e^{-9}$
BERT_BASE	0.17	0.16	[0.006,0.02]	0.002	0.06	0.08	[0.01,0.02]	$4e^{-6}$
BERT_PRO	0.63	0.56	[0.04,0.07]	$1e^{-13}$	0.26	0.32	[0.05,0.07]	$< 2.2e^{-16}$
UmBERTo	0.06	0.05	[0.01,0.02]	$4e^{-6}$	0.009	0.02	[0.004,0.01]	$2e^{-9}$
UmBERTo_BASE	0.12	0.09	[0.02,0.04]	$3e^{-9}$	0.01	0.02	[0.01,0.02]	$9e^{-12}$
UmBERTo_PRO	0.67	0.58	[0.07,0.11]	$5e^{-16}$	0.19	0.28	[0.07,0.11]	$< 2.2e^{-16}$
GilBERTo	0.26	0.25	[-0.006,0.02]	0.30	0.20	0.22	[0.01,0.03]	0.0002
GilBERTo_BASE	0.24	0.24	[-0.006,0.01]	0.44	0.16	0.18	[0.01,0.03]	$3e^{-7}$
GilBERTo_PRO	0.54	0.50	[0.03,0.06]	$3e^{-7}$	0.40	0.45	[0.04,0.07]	$3e^{-10}$
mBERT	0.13	0.14	[-0.004,-0.02]	0.0003	0.12	0.13	[0.003,0.02]	0.003

Table A.9: Results from pairwise t-tests for Italian across the investigated models from Parisi et al. (2020) and Ravasio and Di Perna (2020). O-O refers to object antecedent after object-biased IC verb and O-S to object antecedent after subject-biased IC verb (similarly for subject antecedents S-O and S-S). CI is 95% confidence intervals (where positive is an IC effect). BERT_BASE and BERT_PRO refer to models fine-tuned on baseline data and data with a pro drop process respectively.

	SG	PL	NA
1	3769	977	-
2	-	-	1958
3	2475	1207	-

Table A.10: Breakdown of pronouns removed for English fine-tuning data. Pronoun person and number were determined by annotations in UD data, with NA being pronouns unmarked for number. There were a total of 6871 sentences comprised of 109650 tokens in the training set.

	SG	PL	NA
1	-	56	66
2	-	2	21
3	-	164	774

Table A.11: Breakdown of pronouns removed for Chinese fine-tuning data. Pronoun person and number were determined by annotations in UD data, with NA being pronouns unmarked for number. There were a total of 935 sentences comprised of 108949 characters in the training set.

	SG	PL	NA
1	519	417	-
2	99	7	-
3	3574	944	-

Table A.12: Breakdown of pronouns added for Spanish fine-tuning data. Pronoun person and number were determined by annotations in UD data, with NA being pronouns unmarked for number. There were a total of 4000 sentences comprised of 5559 tokens in the training set.

	SG	PL	NA
1	654	417	-
2	399	94	-
3	2284	679	-

Table A.13: Breakdown of pronouns added for Italian fine-tuning data. Pronoun person and number were determined by annotations in UD data, with NA being pronouns unmarked for number. There were a total of 3798 sentences comprised of 4608 tokens in the training set.

APPENDIX B

APPENDIX FOR AMBIGUOUS RELATIVE CLAUSE ATTACHMENT

B.1 Neural Models and Attachment Preferences

For Section 4.4, English stimuli are given in Table B.1, and Spanish stimuli are given in Table B.2.

B.2 Fine-Grained Attachment Preferences in Neural Models

For Section 4.5, English stimuli are given in Table B.3, and Spanish stimuli are given in Table B.4.

B.3 Interaction between Attachment and Implicit Causality in English

For Section 4.6, the stimuli are given in Table B.5.

B.4 Gender Agreement and Attachment in Spanish

For Section 4.7, the stimuli are given in Table B.6.

item	sentence
1a	Andrew had dinner yesterday with the relative of the teacher that MASK divorced.
2a	The journalist interviewed the coach of the athlete that MASK sick.
3a	The personnel manager was observing the secretary of the clerk that MASK studying.
4a	Julia had spoken to the secretary of the lawyer that MASK on vacation.
5a	My friend met the assistant of the detective that MASK fired.
6a	Charlie met the employee of the ambassador that MASK eating.
7a	Roxanne read the review of the poem that MASK unfinished.
8a	The plumber adjusted the pipe of the sink that MASK cracked.
9a	Mary replaced the wire of the speaker that MASK damaged.
10a	My brother liked listening to the recording of the song that MASK banned.
11a	The chef couldn't find the cover of the pot that MASK clean.
12a	The thief took the key of the car that MASK outside.
13a	The journalist was unable to interview the daughter of the hostage that MASK waiting.
14a	Patricia saw the teacher of the student that MASK in class.
15a	Linda wrote to the manager of the assistant that MASK late.
16a	The hotel director didn't want to see the guide of the tourist that MASK angry.
17a	The receptionist greeted the client of the lawyer that MASK chatting.
18a	Nobody noticed the guard of the ambassador that MASK hiding.
19a	Ivana met the son of the delegate that MASK smoking.
20a	Lisa couldn't find the replacement of the pen that MASK on sale.
21a	The student read the revision of the manuscript that MASK on the test.
22a	The archaeologists finally found the panel of the box that MASK broken.
23a	Harry had inspected the printer of the computer that MASK stolen.
24a	Susan admired the hall of the apartment that MASK painted.
1	Peter was looking at the book of the girl that MASK in the living room.
2	Someone shot the servant of the actress who MASK on the balcony.
3	John met the friend of the teacher who MASK in Germany.
4	A thief was keeping an eye on the case of the tourist that MASK by the mailbox.
5	The tourist photographed the animal of the peasant that MASK by the puddle.
6	The police arrested the sister of the driver who MASK in Melilla.

item	sentence
7	The nurse took the medicine of the patient that MASK by the window.
8	Andrew was speaking with the sibling of the farmer who MASK in Brazil.
9	The servant contemplated the shoe of the guest that MASK close to the fireplace.
10	Yesterday I met with the girlfriend of the town leader who MASK in Australia.
11	The bailiff locked up the horse of the traveler that MASK by the bridge.
12	The old lady was observing the toy of the baby that MASK on the bed.
13	The journalist interviewed the daughter of the captain who MASK near the accident.
14	The peasant was gazing at the bag of the visitor that MASK under a tree.
15	The people watched the box of the soldier that MASK on the platform.
16	Lewis ran over the dog of the fruit dealer that MASK to this district.
17	Andrew had dinner with the relative of the worker who MASK a member of the communist party.
18	George was stroking the cat of the french girl that MASK at the fountain.
19	Martha cheered the brother of the priest who MASK in the school.
20	The detective photographed the container of the student that MASK on the terrace.
21	Mary argued with the cousin of the man who MASK in Argentina.
22	The boys poked fun at the child of the teacher who MASK in the park.
23	My mother argued with the servant of the lady who MASK in the house.
24	This afternoon I saw the son of the doctor who MASK at our home.

Table B.1: Templates for English stimuli for Section 4.4. Item numbers with "a" are adapted from Fernández (2003), and the others are adapted from Cuetos and Mitchell (1988). The MASK was replaced by the model specific MASK token or used as the truncation point. The full stimuli vary the number on the nouns in the complex noun phrase.

item	sentence
1a	Andrés cenó ayer con el relativo del maestro que MASK en el partido comunista..
2a	El periodista entrevistó al entrenador del atleta que MASK en la competición.
3a	El gerente observaba al asistente del trabajador que MASK en la oficina.
4a	Julia había hablado con al asistente del abogado que MASK de vacación.
5a	Mi amigo conoció al ayudante del detective que MASK en la oficina.
6a	Carlos conoció al analista del embajador que MASK en la compañia.
7a	Rosa leyó el verso de la poema que MASK en las últimas páginas de la revista.
8a	El plomero ajustó el tubo de la cocina que MASK en la casa.
9a	María remplazó el cable del radio que MASK en el coche.
10a	A mi hermano le gustaba escuchar el ritmo del canto que MASK en la primera cara del álbum.
11a	El cocinero no pudo encontrar el color del plato que MASK en el aparador.
12a	El ladrón se llevó la llave de la caja que MASK en al armario del pasillo.
13a	El periodista no pudo entrevistar a la hija del prisionero que MASK en el avión.
14a	Patricia vio al profesor del estudiante que MASK en clase.
15a	Linda escribió al director del asistente que MASK en la universidad.
16a	El director del hotel no quiso ver al conductor del turista que MASK en recepción..
17a	La recepcionista saludó al cliente del abogado que MASK en la sala de conferencias.
18a	Nadie vio al guarda del embajador que MASK en la fiesta.
19a	Ivana conoció al hijo del delegado que MASK en el salón.
20a	Lisa no pudo encontrar el componente de la pluma que MASK de oferta.
21a	El estudiante leyó el relato del manuscrito que MASK en el examen.
22a	Los arqueólogos finalmente encontraron el panel de la caja que MASK en el poema.
23a	Enrique había inspeccionado la pieza del computadora que MASK encima del escritorio..
24a	Susana admiró el pasillo del apartamento que MASK cerca del parque.
1	Pedro miraba el libro de la chica que MASK en el salón.
2	Alguien disparó contra el criado de la actriz que MASK en el balcón.
3	Juan conoció al amigo de la maestra que MASK en Alemania.
4	Un ladrón espiaba la maleta del turista que MASK junto al buzón.
5	El turista fotografió el caballo del campesino que MASK junto a la charca.

item	sentence
6	La policía detuvo a la hermana del trabajador que MASK en Melilla.
7	La enfermera apartó la medicina del paciente que MASK junto a la ventana.
8	Andrés estuvo hablando con la relativa del granjero que MASK en Brasil.
9	La sirvienta contemplaba el zapato del invitado que MASK junto a la chimenea.
10	Ayer me encontré con la amiga del concejal que MASK en Australia.
11	El alguacil encerró al caballo del viajero que MASK junto al puente.
12	La anciana observaba el juguete del chico que MASK encima de la cama.
13	El periodista entrevistó a la hija del comandante que MASK el accidente.
14	El campesino contemplaba la maleta del viajero que MASK bajo un árbol.
15	La gente observaba la caja del soldado que MASK en el andén.
16	Luis atropelló al perro del vendedor que viene a MASK barrio.
17	Andrés cenó ayer con la prima del obrero que MASK en partido comunista.
18	Jorge acariciaba al gato de la francesa que MASK en la fuente.
19	Marta saludó al hermano del cura que MASK en el colegio.
20	El detective fotografió la maleta del estudiante que MASK en la terraza.
21	María discutió con la prima del vaquero que MASK en Argentina.
22	Los chicos se burlaron de la chica del maestro que MASK en el parque.
23	Mi madre discutió con el esclavo del rey que MASK de casa.
24	Esta tarde he visto al hijo del doctor que MASK en nuestra casa.

Table B.2: Templates for Spanish stimuli for Section 4.4. Item numbers with "a" are adapted from Fernández (2003), and the others are adapted from Cuetos and Mitchell (1988). The MASK was replaced by the model specific MASK token or used as the truncation point. The full stimuli vary the number on the nouns in the complex noun phrase.

item	exp	sentence
1	A	In the garage we keep the table of wood that MASK carved this Christmas holiday.
2	A	In the garage we keep a table of wood that MASK carved this Christmas holiday.
3	A	To my sister they gave the lamp of material that MASK polished until it looked like marble.
4	A	To my sister they gave a lamp of material that MASK polished until it looked like marble.
5	A	Yesterday they gave me the shirt of fabric that MASK illegally imported.
6	A	Yesterday they gave me a shirt of fabric that MASK illegally imported.
7	A	In the end Tomas brought the blanket of material that MASK very expensive.
8	A	In the end Tomas brought a blanket of material that MASK very expensive.
9	A	Maria made the belt of hide that MASK liked a lot.
10	A	Maria made a belt of hide that MASK liked a lot.
11	A	Finally they placed the bell of medal that MASK brought from the foundry.
12	A	Finally they placed a bell of medal that MASK brought from the foundry.
13	A	Yesterday we ate the cake of grain that MASK sold in the oriental shop.
14	A	Yesterday we ate a cake of grain that MASK sold in the oriental shop.
15	A	The young actress admired the dress of thread that MASK so beautiful.
16	A	The young actress admired a dress of thread that MASK so beautiful.
17	A	John asked for the glass of water that MASK on the table.
18	A	John asked for the clear glass of water that MASK on the table.
19	A	John asked for the glass of clear water that MASK on the table.
20	A	Mary liked the bottle of spirit that MASK in the wine cellar.
21	A	Mary liked the old bottle of spirit that MASK in the wine cellar.
22	A	Mary liked the bottle of old spirit that MASK in the wine cellar.
23	A	Andres picked up the sack of sand that MASK brought from the construction site.
24	A	Andres picked up the brown sack of sand that MASK brought from the construction site.

item	exp	sentence
25	A	Andres picked up the sack of brown sand that MASK brought from the construction site.
26	A	The clerk brought us the package of food that MASK on the counter.
27	A	The clerk brought us the gross package of food that MASK on the counter.
28	A	The clerk brought us the package of gross food that MASK on the counter.
29	A	In the dining-room you will find the basket of fruit that MASK on the table.
30	A	In the dining-room you will find the big basket of fruit that MASK on the table.
31	A	In the dining-room you will find the basket of big fruit that MASK on the table.
32	A	Julia picked up the can of oil that MASK oily.
33	A	Julia picked up the light can of oil that MASK oily.
34	A	Julia picked up the can of light oil that MASK oily.
35	A	My mother didn't see the jar of preserve that MASK crawling with ants.
36	A	My mother didn't see the ruined jar of preserve that MASK crawling with ants.
37	A	My mother didn't see the jar of ruined preserve that MASK crawling with ants.
38	B	The teacher was talking with the relative of the boy who MASK in the hospital.
39	B	The teacher was talking with a relative of the boy who MASK in the hospital.
40	B	The teacher was talking with the relative of a boy who MASK in the hospital.
41	B	The journalist had interviewed the daughter of the captain who MASK in an accident.
42	B	The journalist had interviewed a daughter of the captain who MASK in an accident.
43	B	The journalist had interviewed the daughter of a captain who MASK in an accident.
44	B	Andres had dinner with the relative of the servant who MASK happy last summer.

item	exp	sentence
45	B	Andres had dinner with a relative of the servant who MASK happy last summer.
46	B	Andres had dinner with the relative of a servant who MASK happy last summer.
47	B	This morning I met the mother of the mechanic who MASK in the building where I live.
48	B	This morning I met an mother of the mechanic who MASK in the building where I live.
49	B	This morning I met the mother of a mechanic who MASK in the building where I live.
50	B	The police arrested the cousin of the artist who MASK in Marbella.
51	B	The police arrested a cousin of the artist who MASK in Marbella.
52	B	The police arrested the cousin of a artist who MASK in Marbella.
53	B	The doorman was talking to the sister of the nurse who MASK my mother's friend.
54	B	The doorman was talking to a sister of the nurse who MASK my mother's friend.
55	B	The doorman was talking to the sister of a nurse who MASK my mother's friend.
56	B	Everybody in the office felt sorry about the death of the brother of the director who MASK in the corporation.
57	B	Everybody in the office felt sorry about the death of a brother of the director who MASK in the corporation.
58	B	Everybody in the office felt sorry about the death of the brother of a director who MASK in the corporation.
59	B	All of our friends liked the friend of the exchange student who MASK visiting us.
60	B	All of our friends liked an friend of the exchange student who MASK visiting us.

item	exp	sentence
61	B	All of our friends liked the friend of an exchange student who MASK visiting us.
62	B	We were worried about the mother of the au pair who MASK ill.
63	B	We were worried about a mother of the au pair who MASK ill.
64	B	We were worried about the mother of a au pair who MASK ill.
65	B	The explosion deafened the assistant of the inspector who MASK near the warehouse.
66	B	The explosion deafened an assistant of the inspector who MASK near the warehouse.
67	B	The explosion deafened the assistant of an inspector who MASK near the warehouse.
68	B	The police arrested the driver of the actor who MASK accused of dealing drugs.
69	B	The police arrested a driver of the actor who MASK accused of dealing drugs.
70	B	The police arrested the driver of an actor who MASK accused of dealing drugs.
71	B	Next month they will assign to the foreigner the secretary of the manager who MASK long hours in the office.
72	B	Next month they will assign to the foreigner a secretary of the manager who MASK long hours in the office.
73	B	Next month they will assign to the foreigner the secretary of a manager who MASK long hours in the office.
74	B	Most of the patients like the nurse of the surgeon who MASK in the hospital.
75	B	Most of the patients like a nurse of the surgeon who MASK in the hospital.
76	B	Most of the patients like the nurse of a surgeon who MASK in the hospital.
77	B	I was talking to the pupil of the dress maker who MASK in Paris for a while.
78	B	I was talking to an pupil of the dress maker who MASK in Paris for a while.
79	B	I was talking to the pupil of a dress maker who MASK in Paris for a while.
80	B	Tomorrow I have a date with the advisor of the assistant district attorney that MASK at Mary's party.

item	exp	sentence
81	B	Tomorrow I have a date with an advisor of the assistant district attorney that MASK at Mary's party.
82	B	Tomorrow I have a date with the advisor of an assistant district attorney that MASK at Mary's party.
83	B	I was told that the doctor of the performer who MASK on tv yesterday had been fired.
84	B	I was told that a doctor of the performer who MASK on tv yesterday had been fired.
85	B	I was told that the doctor of a performer who MASK on tv yesterday had been fired.
86	B	Yesterday I saw the consultant of the director who MASK upset because of the pitiful response to the latest sales promotion.
87	B	Yesterday I saw a consultant of the director who MASK upset because of the pitiful response to the latest sales promotion.
88	B	Yesterday I saw the consultant of a director who MASK upset because of the pitiful response to the latest sales promotion.
89	B	During the meeting the chief of protocol tried to talk to the employee of the ambassador who MASK at the party.
90	B	During the meeting the chief of protocol tried to talk to a employee of the ambassador who MASK at the party.
91	B	During the meeting the chief of protocol tried to talk to the employee of an ambassador who MASK at the party.
92	B	The tourists admired the museum of the city that MASK beautiful.
93	B	The tourists admired the small museum of the city that MASK beautiful.
94	B	The tourists admired the museum of the small city that MASK beautiful.
95	B	John smashed the car of the company that MASK hated so much.
96	B	John smashed the new car of the company that MASK hated so much.
97	B	John smashed the car of the new company that MASK hated so much.
98	B	Several men moved the machine of the company that MASK on fire.
99	B	Several men moved the old machine of the company that MASK on fire.

item	exp	sentence
100	B	Several men moved the machine of the old company that MASK on fire.
101	B	The brokers sold the stock of the investment club that MASK losing money.
102	B	The brokers sold the new stock of the investment club that MASK losing money.
103	B	The brokers sold the stock of the new investment club that MASK losing money.
104	B	The governor bought some books for the library of the elementary school that MASK built downtown.
105	B	The governor bought some books for the new library of the elementary school that MASK built downtown.
106	B	The governor bought some books for the library of the new elementary school that MASK built downtown.
107	B	The local newspaper columnist wrote about the representative of the club that MASK found so ridiculous.
108	B	The local newspaper columnist wrote about the popular representative of the club that MASK found so ridiculous.
109	B	The local newspaper columnist wrote about the representative of the popular club that MASK found so ridiculous.
110	B	The pilot was looking at the airport through the side window of the plane that MASK fixed.
111	B	The pilot was looking at the airport through the filthy side window of the plane that MASK fixed.
112	B	The pilot was looking at the airport through the side window of the filthy plane that MASK fixed.
113	B	Birds won't be able to nest in the branch of the tree that MASK cut last year.
114	B	Birds won't be able to nest in the big branch of the tree that MASK cut last year.
115	B	Birds won't be able to nest in the branch of the big tree that MASK cut last year.
116	B	The car stopped in front of the door of the house that MASK damaged.

item	exp	sentence
117	B	The car stopped in front of the main door of the house that MASK damaged.
118	B	The car stopped in front of the door of the main house that MASK damaged.
119	B	The plumber suggested to us to change the valve of the sink that MASK installed last year.
120	B	The plumber suggested to us to change the new valve of the sink that MASK installed last year.
121	B	The plumber suggested to us to change the valve of the new sink that MASK installed last year.
122	B	I really liked the chapter of the book MASK read yesterday.
123	B	I really liked the short chapter of the book MASK read yesterday.
124	B	I really liked the chapter of the short book MASK read yesterday.
125	B	Silvia didn't find the cover of the pot that MASK just cleaned.
126	B	Silvia didn't find the old cover of the pot that MASK just cleaned.
127	B	Silvia didn't find the cover of the old pot that MASK just cleaned.
128	B	We have to paint the bar of the bicycle that MASK fixed yesterday.
129	B	We have to paint the blue bar of the bicycle that MASK fixed yesterday.
130	B	We have to paint the bar of the blue bicycle that MASK fixed yesterday.
131	B	In the meeting they showed us the label of the bottle that MASK designed by the artist.
132	B	In the meeting they showed us the new label of the bottle that MASK designed by the artist.
133	B	In the meeting they showed us the label of the new bottle that MASK designed by the artist.
134	B	The insurance inspector photographed the engine of the boat that MASK covered with water.
135	B	The insurance inspector photographed the damaged engine of the boat that MASK covered with water.
136	B	The insurance inspector photographed the engine of the damaged boat that MASK covered with water.
137	B	I was surprised by the sketch of the sculpture that MASK in the town hall.

item	exp	sentence
138	B	I was surprised by the odd sketch of the sculpture that MASK in the town hall.
139	B	I was surprised by the sketch of the odd sculpture that MASK in the town hall.
140	B	The designer agreed to show us the sketch of the house that MASK finished before the end of the summer.
141	B	The designer agreed to show us the new sketch of the house that MASK finished before the end of the summer.
142	B	The designer agreed to show us the sketch of the new house that MASK finished before the end of the summer.
143	B	The architect exhibited the drawing of the building that MASK commissioned.
144	B	The architect exhibited the plain drawing of the building that MASK commissioned.
145	B	The architect exhibited the drawing of the plain building that MASK commissioned.
146	B	Charles liked the portrait of the woman that MASK at your house.
147	B	Charles liked the sad portrait of the woman that MASK at your house.
148	B	Charles liked the portrait of the sad woman that MASK at your house.
149	B	All the newspapers published the photograph of the boy that MASK liked so much.
150	B	All the newspapers published the small photograph of the boy that MASK liked so much.
151	B	All the newspapers published the photograph of the small boy that MASK liked so much.
152	B	Sara did the painting of the cave that MASK so gloomy.
153	B	Sara did the famous painting of the cave that MASK so gloomy.
154	B	Sara did the painting of the famous cave that MASK so gloomy.
155	B	The collector lost the picture of the house that MASK so dark.
156	B	The collector lost the big picture of the house that MASK so dark.
157	B	The collector lost the picture of the big house that MASK so dark.

item	exp	sentence
158	B	Suzanne sold the painting of the beach that MASK liked by her friends
159	B	Suzanne sold the small painting of the beach that MASK liked by her friends
160	B	Suzanne sold the painting of the small beach that MASK liked by her friends
161	B	The critics judged very harshly the print of the park that MASK attracting the most visitors.
162	B	The critics judged very harshly the foreign print of the park that MASK attracting the most visitors.
163	B	The critics judged very harshly the print of the foreign park that MASK attracting the most visitors.
164	B	The professor read the book of the student that MASK in the dining-room.
165	B	The professor read the new book of the student that MASK in the dining-room.
166	B	The professor read the book of the new student that MASK in the dining-room.
167	B	The inspector observed the case of the traveler that MASK in the station.
168	B	The inspector observed the suspicious case of the traveler that MASK in the station.
169	B	The inspector observed the case of the suspicious traveler that MASK in the station.
170	B	The mechanics were modifying the car of the driver that MASK in the race.
171	B	The mechanics were modifying the dangerous car of the driver that MASK in the race.
172	B	The mechanics were modifying the car of the dangerous driver that MASK in the race.
173	B	The dressmaker was sewing the dress of the girl that MASK on the floor.
174	B	The dressmaker was sewing the dirty dress of the girl that MASK on the floor.
175	B	The dressmaker was sewing the dress of the dirty girl that MASK on the floor.
176	B	We borrowed the car of the neighbor that MASK nearby.
177	B	We borrowed the old car of the neighbor that MASK nearby.
178	B	We borrowed the car of the old neighbor that MASK nearby.
179	B	I had to borrow the computer of the secretary that MASK in the office close to mine.

item	exp	sentence
180	B	I had to borrow the new computer of the secretary that MASK in the office close to mine.
181	B	I had to borrow the computer of the new secretary that MASK in the office close to mine.
182	C	The count ordered the sandwich with the side that MASK prepared especially well.
183	C	The count ordered a sandwich with the side that MASK prepared especially well.
184	C	The count ordered the sandwich with a side that MASK prepared especially well.
185	C	Laura lost the book with the label that MASK given to her.
186	C	Laura lost a book with the label that MASK given to her.
187	C	Laura lost the book with a label that MASK given to her.
188	C	Marta was wearing the hat with the dress that MASK worn in summer.
189	C	Marta was wearing a hat with the dress that MASK worn in summer.
190	C	Marta was wearing the hat with a dress that MASK worn in summer.
191	C	It was agreed to move the computer with the screen that MASK brought recently to another building.
192	C	It was agreed to move a computer with the screen that MASK brought recently to another building.
193	C	It was agreed to move the computer with a screen that MASK brought recently to another building.
194	C	I wanted to take the radio with the speaker that MASK bought for a very low price.
195	C	I wanted to take a radio with the speaker that MASK bought for a very low price.
196	C	I wanted to take the radio with a speaker that MASK bought for a very low price.
197	C	On the shelf I keep the box with the cover that MASK varnished.
198	C	On the shelf I keep a box with the cover that MASK varnished.

item	exp	sentence
199	C	On the shelf I keep the box with a cover that MASK varnished.
200	C	The captain authorized the departure of the ship with the pole that MASK repaired.
201	C	The captain authorized the departure of a ship with the pole that MASK repaired.
202	C	The captain authorized the departure of the ship with a pole that MASK repaired.
203	C	Yesterday I dropped the plate with the candle that MASK made in Barcelona.
204	C	Yesterday I dropped a plate with the candle that MASK made in Barcelona.
205	C	Yesterday I dropped the plate with a candle that MASK made in Barcelona.
206	C	The millionaire was shown the house with the pool that MASK as big as half a football field.
207	C	The millionaire was shown a house with the pool that MASK as big as half a football field.
208	C	The millionaire was shown the house with a pool that MASK as big as half a football field.

Table B.3: Templates for the English stimuli for Section 4.5 and adapted from Gilboy et al. (1995). The MASK was replaced by the model specific MASK token or used as the truncation point. The full stimuli vary the number on the nouns in the complex noun phrase.

item	exp	sentence
1	A	En el garaje guardamos la mesa de material que MASK estas Navidades.
2	A	En el garaje guardamos una mesa de material que MASK estas Navidades.
3	A	A mi hermana le regalaron la vela de mineral que MASK en la bodega.
4	A	A mi hermana le regalaron una vela de mineral que MASK en la bodega.
5	A	Ayer me regalaron la camisa de tela que MASK de contrabando.
6	A	Ayer me regalaron una camisa de tela que MASK de contrabando.
7	A	Al final Tomás compró la manta de tela que MASK muy cara.
8	A	Al final Tomás compró una manta de tela que MASK muy cara.
9	A	María hizo la bolsa de piel que MASK mucho.
10	A	María hizo una bolsa de piel que MASK mucho.
11	A	Por fin colocaron la campana de metal que MASK de la fundición.
12	A	Por fin colocaron una campana de metal que MASK de la fundición.
13	A	Ayer nos comimos el pastel de grano que MASK en la tienda de productos orientales.
14	A	Ayer nos comimos un pastel de grano que MASK en la tienda de productos orientales.
15	A	La joven actriz admiraba el vestido de fibra que MASK muy bonita.
16	A	La joven actriz admiraba un vestido de fibra que MASK muy bonita.
17	A	Juan pidió el vaso de jugo que MASK encima de la mesa.
18	A	Juan pidió el vaso transparente de jugo que MASK encima de la mesa.
19	A	Juan pidió el vaso de jugo transparente que MASK encima de la mesa.
20	A	A María le gustaba la botella de vino que MASK en la bodega.
21	A	A María le gustaba la botella vieja de vino que MASK en la bodega.
22	A	A María le gustaba la botella de vino viejo que MASK en la bodega.
23	A	Pedro se llevó la copa de dulce que MASK al suelo.
24	A	Pedro se llevó la copa blanca de dulce que MASK al suelo.
25	A	Pedro se llevó la copa de dulce blanca que MASK al suelo.
26	A	Andrés recogió el saco de arena que MASK de la obra.
27	A	Andrés recogió el saco amarillo de arena que MASK de la obra.
28	A	Andrés recogió el saco de arena amarilla que MASK de la obra.

item	exp	sentence
29	A	El vendedor nos trajo el paquete de carne que MASK en el mostrador.
30	A	El vendedor nos trajo el paquete nuevo de carne que MASK en el mostrador.
31	A	El vendedor nos trajo el paquete de carne nuevo que MASK en el mostrador.
32	A	Cuando vamos de camping llevamos la maleta de ropa que MASK en verano.
33	A	Cuando vamos de camping llevamos la maleta vieja de ropa que MASK en verano.
34	A	Cuando vamos de camping llevamos la maleta de ropa vieja que MASK en verano.
35	A	En el comedor encontrarás el caso de manzana que MASK sobra la mesa.
36	A	En el comedor encontrarás el caso grande de manzana que MASK sobra la mesa.
37	A	En el comedor encontrarás el caso de manzana grande que MASK sobra la mesa.
38	A	Julia recogió el paquete de color que MASK sobra la mesa.
39	A	Julia recogió el paquete claro de color que MASK sobra la mesa.
40	A	Julia recogió el paquete de color claro que MASK sobra la mesa.
41	A	Mi madre no vió la copa de jugo que MASK de hormigas.
42	A	Mi madre no vió la copa rota de jugo que MASK de hormigas.
43	A	Mi madre no vió la copa de jugo roto que MASK de hormigas.
44	B	La maestra estuvo hablando con la pariente del chico que MASK en el hospital.
45	B	La maestra estuvo hablando con una pariente del chico que MASK en el hospital.
46	B	La maestra estuvo hablando con la pariente de un chico que MASK en el hospital.
47	B	Los periodistas entrevistaron a la hija del maestro que MASK un accidente.
48	B	Los periodistas entrevistaron a una hija del coronel que MASK un accidente.
49	B	Los periodistas entrevistaron a la hija de un coronel que MASK un accidente.
50	B	Andrés estuvo cenando con la sobrina del portero que MASK el verano pasado.
51	B	Andrés estuvo cenando con una sobrina del portero que MASK el verano pasado.

item	exp	sentence
52	B	Andrés estuvo cenando con la sobrina de un portero que MASK el verano pasado.
53	B	Esta mañana me econtré con la tía del mecánico que MASK en mi bloque de pisos.
54	B	Esta mañana me econtré con una tía del mecánico que MASK en mi bloque de pisos.
55	B	Esta mañana me econtré con la tía de un mecánico que MASK en mi bloque de pisos.
56	B	La policía detuvo a la prima del pintor que MASK en Marbella.
57	B	La policía detuvo a una prima del pintor que MASK en Marbella.
58	B	La policía detuvo a la prima de un pintor que MASK en Marbella.
59	B	El portero conversaba con la hermana de la enfermera que MASK amiga de mi madre.
60	B	El portero conversaba con una hermana de la enfermera que MASK amiga de mi madre.
61	B	El portero conversaba con la hermana de una enfermera que MASK amiga de mi madre.
62	B	Todos en la oficina sintieron la muerta de la hermana del directivo que MASK trabajado tanto tiempo en la empresa.
63	B	Todos en la oficina sintieron la muerta de una hermana del directivo que MASK trabajado tanto tiempo en la empresa.
64	B	Todos en la oficina sintieron la muerta de la hermana de un directivo que MASK trabajado tanto tiempo en la empresa.
65	B	A todos nuestros amigos les gustaba el hermano del estudiante extranjero que MASK el verano pasado.
66	B	A todos nuestros amigos les gustaba un hermano del estudiante extranjero que MASK el verano pasado.
67	B	A todos nuestros amigos les gustaba el hermano de un estudiante extranjero que MASK el verano pasado.
68	B	Estábamos preocupados por la madre de la muchacha que MASK al hospital.

item	exp	sentence
69	B	Estábamos preocupados por una madre de la muchacha que MASK al hospital.
70	B	Estábamos preocupados por la madre de una muchacha que MASK al hospital.
71	B	La explosión ensordeció al ayudante del ministro que MASK junto al almaén.
72	B	La explosión ensordeció a un ayudante del ministro que MASK junto al almaén.
73	B	La explosión ensordeció al ayudante de un ministro que MASK junto al almaén.
74	B	La policía también detuvo al ayudante del actor que MASK en la empresa.
75	B	La policía también detuvo a un ayudante del actor que MASK en la empresa.
76	B	La policía también detuvo al ayudante de un actor que MASK en la empresa.
77	B	El mes que viene envían al extranjero al asistente del director que MASK un montón de horas en la oficina.
78	B	El mes que viene envían al extranjero a un asistente del director que MASK un montón de horas en la oficina.
79	B	El mes que viene envían al extranjero al asistente de un director que MASK un montón de horas en la oficina.
80	B	A la mayoría de los enfermos les gustaba la enfermera del doctor que MASK de empezar a trabajar en el hospital.
81	B	A la mayoría de los enfermos les gustaba una enfermera del doctor que MASK de empezar a trabajar en el hospital.
82	B	A la mayoría de los enfermos les gustaba la enfermera de un doctor que MASK de empezar a trabajar en el hospital.
83	B	Estuve conversando con el alumno del artista que MASK en Paris durante un tiempo.
84	B	Estuve conversando con un alumno del artista que MASK en Paris durante un tiempo.
85	B	Estuve conversando con el alumno de un artista que MASK en Paris durante un tiempo.
86	B	Mañana tengo una cita con el asesor del fiscal que MASK en la fiesta de María.
87	B	Mañana tengo una cita con un asesor del fiscal que MASK en la fiesta de María.

item	exp	sentence
88	B	Mañana tengo una cita con el asesor de un fiscal que MASK en la fiesta de María.
89	B	Me dijeron que el consejero de la adolescente que MASK en la televisión.
90	B	Me dijeron que un consejero de la adolescente que MASK en la televisión.
91	B	Me dijeron que el consejero de una adolescente que MASK en la televisión.
92	B	Ayer ví al consejero del director que MASK en la oficina.
93	B	Ayer ví a un consejero del director que MASK en la oficina.
94	B	Ayer ví al consejero de un director que MASK en la oficina.
95	B	Durante la reunión el jefe de protocolo intentó hablar con la analista del embajador que MASK en la fiesta.
96	B	Durante la reunión el jefe de protocolo intentó hablar con una analista del embajador que MASK en la fiesta.
97	B	Durante la reunión el jefe de protocolo intentó hablar con la analista de un embajador que MASK en la fiesta.
98	B	Los turistas admiraban el museo de la ciudad que MASK en agosto.
99	B	Los turistas admiraban el museo grande de la ciudad que MASK en agosto.
100	B	Los turistas admiraban el museo de la ciudad grande que MASK en agosto.
101	B	Juan se estrelló con el coche de la compañía que MASK ayer.
102	B	Juan se estrelló con el coche de la compañía que MASK ayer.
103	B	Juan se estrelló con el coche de la compañía que MASK ayer.
104	B	Varios hombres trasladaron el aparato de la tela que se MASK incendiado.
105	B	Varios hombres trasladaron el aparato viejo de la tela que se MASK incendiado.
106	B	Varios hombres trasladaron el aparato de la tela vieja que se MASK incendiado.
107	B	Los consultores vendieron la demanda de la empresa que MASK bajando la cotización.
108	B	Los consultores vendieron la demanda nueva de la empresa que MASK bajando la cotización.
109	B	Los consultores vendieron la demanda de la empresa nueva que MASK bajando la cotización.

item	exp	sentence
110	B	El gobernador compró libros para la biblioteca de la escuela que MASK de construir.
111	B	El gobernador compró libros para la biblioteca de la escuela que MASK de construir.
112	B	El gobernador compró libros para la biblioteca de la escuela que MASK de construir.
113	B	El columnista del periódico escribió sobre la mascota de la asociación que MASK en agosto.
114	B	El columnista del periódico escribió sobre la mascota de la asociación que MASK en agosto.
115	B	El columnista del periódico escribió sobre la mascota de la asociación que MASK en agosto.
116	B	El piloto contemplaba el aeropuerto desde la ventana del barco que MASK limpiando.
117	B	El piloto contemplaba el aeropuerto desde la ventana sucia del barco que MASK limpiando.
118	B	El piloto contemplaba el aeropuerto desde la ventana del barco sucio que MASK limpiando.
119	B	Los ájaros no podán anidar en la rama del arbusto que MASK el ño pasado.
120	B	Los ájaros no podán anidar en la rama grande del arbusto que MASK el ño pasado.
121	B	Los ájaros no podán anidar en la rama del arbusto grande que MASK el ño pasado.
122	B	El coche se detuvo ante la puerta de la casa que MASK claros signos de deterioro.
123	B	El coche se detuvo ante la puerta blanca de la casa que MASK claros signos de deterioro.
124	B	El coche se detuvo ante la puerta de la casa blanca que MASK claros signos de deterioro.

item	exp	sentence
125	B	El fontanero nos recomendó cambiar la llave de la cocina que MASK el ño pasado.
126	B	El fontanero nos recomendó cambiar la llave nueva de la cocina que MASK el ño pasado.
127	B	El fontanero nos recomendó cambiar la llave de la cocina nueva que MASK el ño pasado.
128	B	Me gustó mucho el verso del libro que MASK ayer.
129	B	Me gustó mucho el verso breve del libro que MASK ayer.
130	B	Me gustó mucho el verso del libro breve que MASK ayer.
131	B	Silvia no econtraba la marca del contenedor que MASK de limpiar.
132	B	Silvia no econtraba la marca vieja del contenedor que MASK de limpiar.
133	B	Silvia no econtraba la marca del contenedor viejo que MASK de limpiar.
134	B	Tenemos que pintar la campana de la bicicleta que MASK ayer.
135	B	Tenemos que pintar la campana azul de la bicicleta que MASK ayer.
136	B	Tenemos que pintar la campana de la bicicleta azul que MASK ayer.
137	B	En la reunión nos mostraron la etiqueta de la botella que MASK ayer.
138	B	En la reunión nos mostraron la etiqueta nueva de la botella que MASK ayer.
139	B	En la reunión nos mostraron la etiqueta de la botella nueva que MASK ayer.
140	B	El inspector de seguros fotografió el motor del barco que MASK de agua.
141	B	El inspector de seguros fotografió el motor roto del barco que MASK de agua.
142	B	El inspector de seguros fotografió el motor del barco roto que MASK de agua.
143	B	Me sorprendió el dibujo de la escultura que MASK en el ayuntamiento.
144	B	Me sorprendió el dibujo desconocido de la escultura que MASK en el ayuntamiento.
145	B	Me sorprendió el dibujo de la escultura desconocida que MASK en el ayuntamiento.
146	B	El diseñador accedió a mostrarnos el esquema de la casa que MASK acabar antes del fin del verano.
147	B	El diseñador accedió a mostrarnos el esquema nuevo de la casa que MASK acabar antes del fin del verano.

item	exp	sentence
148	B	El diseñador accedió a mostrarnos el esquema de la casa nueva que MASK acabar antes del fin del verano.
149	B	El arquitecto exhibió el dibujo del edificio que MASK sobre la mesa.
150	B	El arquitecto exhibió el dibujo sencillo del edificio que MASK sobre la mesa.
151	B	El arquitecto exhibió el dibujo del edificio sencillo que MASK sobre la mesa.
152	B	A Carlos le gustó el retrato de la mujer que MASK en tu casa.
153	B	A Carlos le gustó el retrato triste de la mujer que MASK en tu casa.
154	B	A Carlos le gustó el retrato de la mujer triste que MASK en tu casa.
155	B	En todos los periódicos publicaron la foto del chico que MASK mucho.
156	B	En todos los periódicos publicaron la foto grandes del chico que MASK mucho.
157	B	En todos los periódicos publicaron la foto del chico grande que MASK mucho.
158	B	Sara pintó el cuadro de la cueva que MASK cerca de la mesa.
159	B	Sara pintó el cuadro famoso de la cueva que MASK cerca de la mesa.
160	B	Sara pintó el cuadro de la cueva famosa que MASK cerca de la mesa.
161	B	El coleccionista perdió la manta de la casa que MASK cerca de la mesa.
162	B	El coleccionista perdió la manta enorme de la casa que MASK cerca de la mesa.
163	B	El coleccionista perdió la manta de la casa enorme que MASK cerca de la mesa.
164	B	Susana vendió la pintura de la playa que MASK cera de sus amigos.
165	B	Susana vendió la pintura grande de la playa que MASK cera de sus amigos.
166	B	Susana vendió la pintura de la playa grande que MASK cera de sus amigos.
167	B	La crítica juzgó duramente la pintura del parque que MASK ayer.
168	B	La crítica juzgó duramente la pintura extranjera del parque que MASK ayer.
169	B	La crítica juzgó duramente la pintura del parque extranjero que MASK ayer.
170	B	El profesor leía el libro del estudiante que MASK en el salón.
171	B	El profesor leía el libro nuevo del estudiante que MASK en el salón.
172	B	El profesor leía el libro del estudiante nuevo que MASK en el salón.
173	B	El revisor observaba la maleta del viajero que MASK en la estación.
174	B	El revisor observaba la maleta vieja del viajero que MASK en la estación.

item	exp	sentence
175	B	El revisor observaba la maleta del viajero viejo que MASK en la estación.
176	B	Los mecánicos revisaban el coche del piloto que MASK en la carrera.
177	B	Los mecánicos revisaban el coche peligroso del piloto que MASK en la carrera.
178	B	Los mecánicos revisaban el coche del piloto peligroso que MASK en la carrera.
179	B	La modista cosía el vestido de la chica que MASK en el suelo.
180	B	La modista cosía el vestido sucio de la chica que MASK en el suelo.
181	B	La modista cosía el vestido de la chica sucia que MASK en el suelo.
182	B	Pedimos prestado el coche del vecino que MASK por allí cerca.
183	B	Pedimos prestado el coche viejo del vecino que MASK por allí cerca.
184	B	Pedimos prestado el coche del vecino viejo que MASK por allí cerca.
185	B	Tuve que pedir prestado el ordenador del ministro que MASK en el despacho al lado del mío.
186	B	Tuve que pedir prestado el ordenador nuevo del ministro que MASK en el despacho al lado del mío.
187	B	Tuve que pedir prestado el ordenador del ministro nuevo que MASK en el despacho al lado del mío.
188	C	El conde pidió la comida con el tomate que MASK especialmente bien.
189	C	El conde pidió una comida con el tomate que MASK especialmente bien.
190	C	El conde pidió la comida con un tomate que MASK especialmente bien.
191	C	Laura perdió el libro con la cinta que MASK en el salón.
192	C	Laura perdió un libro con la cinta que MASK en el salón.
193	C	Laura perdió el libro con una cinta que MASK en el salón.
194	C	Marta se puso el sombrero con la cuerda que MASK en verano.
195	C	Marta se puso un sombrero con la cuerda que MASK en verano.
196	C	Marta se puso el sombrero con una cuerda que MASK en verano.
197	C	Se decidió trasladar el ordenador con la pantalla que MASK a otro edificio.
198	C	Se decidió trasladar un ordenador con la pantalla que MASK a otro edificio.
199	C	Se decidió trasladar el ordenador con una pantalla que MASK a otro edificio.
200	C	Quise llevarme la radio con el cable que MASK por muy poco precio.
201	C	Quise llevarme una radio con el cable que MASK por muy poco precio.

item	exp	sentence
202	C	Quise llevarme la radio con un cable que MASK por muy poco precio.
203	C	En la estantería guardo la caja con la flor que MASK en el salón.
204	C	En la estantería guardo una caja con la flor que MASK en el salón.
205	C	En la estantería guardo la caja con una flor que MASK en el salón.
206	C	El capitán autorizó la salida del buque con el pilar que MASK.
207	C	El capitán autorizó la salida de un buque con el pilar que MASK.
208	C	El capitán autorizó la salida del buque con un pilar que MASK.
209	C	Llevé al joyero la banda con el diamante que MASK en Colombia.
210	C	Llevé al joyero una banda con el diamante que MASK en Colombia.
211	C	Llevé al joyero la banda con un diamante que MASK en Colombia.
212	C	Al millonario se le mostró la casa con la piscina que MASK en Colombia
213	C	Al millonario se le mostró una casa con la piscina que MASK en Colombia
214	C	Al millonario se le mostró la casa con una piscina que MASK en Colombia

Table B.4: Templates for the Spanish stimuli for Section 4.5 and adapted from Gilboy et al. (1995). The MASK was replaced by the model specific MASK token or used as the truncation point. The full stimuli vary the number on the nouns in the complex noun phrase.

item	hasIC	sentence
1	1	The woman scolded the chefs of the aristocrat who MASK routinely letting food go to waste.
1	0	The woman studied with the chefs of the aristocrat who MASK routinely letting food go to waste.
2	1	The man stared at the teachers of the second grader who MASK definitely smartest in the school.
2	0	The man lived next to the teachers of the second grader who MASK definitely smartest in the school.
3	1	The woman assisted the maids of the executive who MASK regularly late to work.
3	0	The woman joked with the maids of the executive who MASK regularly late to work.
4	1	The man trusted the captains of the sailor who MASK consistently weathered big storms.
4	0	The man stood near the captains of the sailor who MASK consistently weathered big storms.
5	1	The woman corrected the secretaries of the lawyer who MASK occasionally made small mistakes.
5	0	The woman gossiped with the secretaries of the lawyer who MASK occasionally made small mistakes.
6	1	The man comforted the leaders of the activist who MASK deeply disappointed by the court's decision.
6	0	The man greeted the leaders of the activist who MASK deeply disappointed by the court's decision.
7	1	The woman envies the managers of the cashier who MASK supposedly received a huge raise.
7	0	The woman knows the managers of the cashier who MASK supposedly received a huge raise.
8	1	The man valued the daughters of the shopkeeper who MASK usually willing to spot him a few dollars.

item	hasIC	sentence
8	0	The man recognized the daughters of the shopkeeper who MASK usually willing to spot him a few dollars.
9	1	The woman fears the uncles of the toddler who MASK often heard yelling and screaming.
9	0	The woman jogs with the uncles of the toddler who MASK often heard yelling and screaming.
10	1	The man noticed the representatives of the employee who MASK always wearing safety goggles.
10	0	The man resembled the representatives of the employee who MASK always wearing safety goggles.
11	1	The woman praised the gardeners of the millionaire who MASK recently installed a solar powered sprinkler.
11	0	The woman met the gardeners of the millionaire who MASK recently installed a solar powered sprinkler.
12	1	The man hates the cousins of the accountant who MASK forever telling the same tasteless jokes.
12	0	The man carpools with the cousins of the accountant who MASK forever telling the same tasteless jokes.
13	1	The woman blamed the nieces of the florist who MASK repeatedly ruined expensive orchids.
13	0	The woman waited with the nieces of the florist who MASK repeatedly ruined expensive orchids.
14	1	The man helped the brothers of the athlete who MASK perpetually failing math class.
14	0	The man ran into the brothers of the athlete who MASK perpetually failing math class.
15	1	The woman reproached the doctors of the supermodel who MASK adamantly in favor of plastic surgery.
15	0	The woman worked with the doctors of the supermodel who MASK adamantly in favor of plastic surgery.

item	hasIC	sentence
16	1	The man pacified the associates of the businessman who MASK nearly bankrupted by the new tax policy.
16	0	The man visited the associates of the businessman who MASK nearly bankrupted by the new tax policy.
17	1	The woman detests the children of the musician who MASK generally arrogant and rude.
17	0	The woman babysits the children of the musician who MASK generally arrogant and rude.
18	1	The man thanked the servants of the dictator who MASK lately been helping the poor.
18	0	The man talked to the servants of the dictator who MASK lately been helping the poor.
19	1	The woman congratulated the bodyguards of the celebrity who MASK constantly fighting off the paparazzi.
19	0	The woman chatted with the bodyguards of the celebrity who MASK constantly fighting off the paparazzi.
20	1	The man mocked the fans of the singer who MASK continually stage diving and getting hurt.
20	0	The man counted the fans of the singer who MASK continually stage diving and getting hurt.

Table B.5: Templates for the stimuli for Section 4.6 and adapted from Rohde et al. (2011). The MASK was replaced by the model specific MASK token or used as the truncation point. Whether the sentence contains an object-biased IC verb is marked by hasIC. The full stimuli vary the number on the nouns in the complex noun phrase.

item	sentence
1	Alguien disparó contra el criado del actor que estaba MASK.
2	Pedro conoció al amigo del maestro que estuvo MASK por el Ministerio.
3	La policía detuvo al hermano del trabajador que estuvo MASK de hurto.
4	Un alumno apedreó al amigo del abogado que estuvo como MASK en el Parlamento.
5	Amelia se fotografió con el novio del cantante que estuvo MASK con los periodistas.
6	El periodista entrevistó al hijo del diputado que se quedó MASK.
7	Andrés cenó ayer con el sobrino del conserje que trabajó de MASK en su empresa.
8	María salió al cine con el hijo del obrero que estaba MASK.
9	Los chicos se burlaron del sobrino del maestro que estaba MASK en el parque.
10	Mi madre discutió con el sirviente del rey que estuvo MASK la semana pasada.
11	La policía detuvo al criado del emperador que estuvo MASK antes por escandalo.
12	Esta tarde he visto al ayudante del doctor que había sido MASK el año pasado.
13	Los chicos se reían con el abuelo del chico que estaba MASK con un traje nuevo.
13	El cartero se acercó al secretario del gerente que estaba MASK.
15	El turista hizo un dibujo del nieto del campesino que estaba MASK.
16	La enfermera tropezó con el visitante del paciente que estaba MASK.
17	El alguacil encerró al hijo del inmigrante que estaba MASK pidiendo limosna.
18	Los rumores acusaban al hermano del propietario que estaba MASK.
19	El detective hizo una foto al amigo del estudiante que estaba MASK.
20	María saludó al hermano del ministro que estaba MASK de volver a su pueblo.
21	Ayer me encontré con el amigo del jefe que fue MASK de nuestra empresa.
22	El periodista entrevistó al secretario del gerente que estaba MASK.
23	Pedro se divertía con el hermano del chico que estaba MASK.
24	La explosión alcanzó al ayudante del diputado que fue ascendido por sus MASK.

Table B.6: Template for the stimuli for Section 4.7 and adapted from Carreiras and Clifton (1993). The MASK was replaced by the model specific MASK token or used as the truncation point. The full stimuli vary the gender on the nouns in the complex noun phrase.

APPENDIX C

APPENDIX FOR PRINCIPLE B AND COREFERENCE

C.1 Principle B as a Constraint on Accessibility: 2 NPs

For Section 5.4, the stimuli are given in Table C.1.

C.2 Principle B as a Constraint on Accessibility: 3 NPs

For Section 5.5, the stimuli are given in Table C.2.

C.3 Predictive Processing with Cataphora

For Section 5.6, the stimuli are given in Table C.3.

C.4 Interaction between Principle B and Predictive Processing

For Section 5.7, the stimuli are given in Table C.4.

item	sentence
1	The man thought that the waiter would praise MASK co-worker for the success of the event.
2	The man worried that the tailor would criticize MASK assistant harshly for the lack of organization at the fashion show.
3	The man revealed that the producer had doubted MASK ability even after several successful performances of the show.
4	The man said that the monk had hidden MASK belief effectively from the persistent agents of the secret police.
5	The man believed that the stock broker had deceived MASK boss repeatedly about the extent of the illegal activity.
6	The man said that the football player had embarrassed MASK friends repeatedly in front of the visiting college recruiters.
7	The man worried that the drug addict would resent MASK body fairly when the withdrawal symptoms became unbearable.
8	The man recalled that the police officer had contradicted MASK lawyer completely during the intense cross examination.
9	The man insisted that the building contractor should familiarize MASK coworkers thoroughly with every detail of the plans.
10	The man denied that the football coach had entertained MASK friend completely by making fun of the students.
11	The man worried that the air traffic controller would blame MASK error entirely for the accident during the emergency landing.
12	The man discovered that the analyst had mocked MASK coworker completely for singing karaoke after drinking too much.
13	The man dreamed that the clown had dressed MASK helper horribly in a frilly costume and an enormous hat.
14	The man believed that the movie director would introduce MASK protege eagerly to the most influential people in the room.
15	The man said that the farmer had reminded MASK colleagues frequently about the dangers of pesticides.

item	sentence
16	The man expected that the pirate would blame MASK foolishness when the curse of the secret treasure was unleashed.
17	The man ensured that the sound engineer had prepared MASK crew thoroughly for any potential mishaps during the performance.
18	The man feared that the choir boy would disappoint MASK audience eventually by hitting a false note in a difficult part.
19	The man claimed that the wrestling coach had pushed MASK team constantly for the sake of improving performance.
20	The man hoped that the sports fan would nominate MASK cashier promptly for the quality service award.
21	The man knew that the bartender would protect MASK friend despite public pressure to expose philandering politicians.
22	The man hoped that the hip hop dancer would teach MASK students amazing moves for any kind of music.
23	The man expected that the pilot would congratulate MASK crew fully for having saved so many lives.
24	The man wished that the garden worker had asked MASK accounant beforehand whether it was worth the money.
25	The man said that the hockey player had defended MASK friend fully despite the constant criticism from the media.
26	The man thought that the director had considered MASK work extremely important to the success of the film.
27	The man claimed that the manager had undermined MASK coworkers constantly in an attempt to climb the corporate ladder.
28	The man announced that the programmer had embarrassed MASK colleagues deeply by failing to notice the glitch ahead of time.
29	The man claimed that the personal trainer had pushed MASK body reasonably hard during the intense training sessions.
30	The man emphasized that the drummer should observe MASK instrument carefully during rehearsals to keep a more consistent rhythm.

item	sentence
31	The man thought that the store owner should thank MASK boss openly for helping to make the necessary changes and cutbacks.
32	The man suspected that the priest would doubt MASK beliefs privately while defending the church publicly.
33	The man hoped that the soldier could trust MASK process completely for survival in the ever worsening situation.
34	The man knew that the union worker would defend MASK strategy fully in the face of political and social pressures.
35	The man lamented that the congressman had humiliated MASK ideas intentionally at the international summit last year.
36	The man said that the consultant had prepared MASK notes sufficiently to make a statement at the press meeting.
37	The man recalled that the physics major would challenge MASK students constantly with difficult questions and criticisms in every class.
38	The man feared that the boss would betray MASK friends eagerly during the interrogation to escape a lengthy prison sentence.
39	The man hoped that the gun advocate would correct MASK thoughts about public opinion on the controversial topic.
40	The man predicted that the metal worker would confront MASK boss daily until salaries were increased for all employees.
41	The man hoped that the school principal would defend MASK ideals fully throughout the controversy over the new legislation.
42	The man supposed that the clown could amuse MASK clients nightly by doing tricks with fire.
43	The man complained that the doctor had pushed MASK patients constantly to lose a few pounds and eat more vegetables.
44	The man said that the fisherman had entertained MASK family publicly by singing lively songs and dancing.
45	The man feared that the market analyst would contradict MASK interests without considering the effects of the worsening housing market.

item	sentence
46	The man predicted that the worker would blame MASK boss alone for the sloppy bookkeeping that led to the investigation.
47	The man claimed that the drummer had pressured MASK group continually to delay the start of quiet hours each night.
48	The man warned that the personal trainer had injured MASK body often by not allowing enough time to stretch before workouts.
49	The man declared that the salesman had failed MASK clients horribly during the big annual sale last weekend.
50	The man heard that the prophet had designated MASK friend officially as the next tribe leader before the election was held.
51	The man mentioned that the comedian had reminded MASK coworker repeatedly to arrive on time for the cover shoot.
52	The man expected that the patriarch would nominate MASK friend readily to marry the butcher's daughter.
53	The man dreamed that the wizard would poison MASK lover secretly on the night of the full moon.
54	The man reported that the professor had introduced MASK clients nicely to the board members at the reception.
55	The man suggested that the engineer should email MASK boss promptly with a summary of what had been done so far.
56	The man remembered that the rock star had blamed MASK manager fairly for the damage to the sound equipment during the show.
57	The man argued that the anarchist would undermine MASK base without helping the cause or gaining any public sympathy.
58	The man appreciated that the construction worker had taught MASK friend patiently about how to make something from scratch.
59	The man assumed that the art critic would promote MASK work openly rather than support stale traditions.
60	The man remembered that the district attorney had congratulated MASK work publicly for helping to raise votes for the new law.

item	sentence

Table C.1: Templates for the stimuli for Section 5.4 and adpated from Chow et al. (2014). The MASK was replaced by the model specific MASK token or used the the truncation point. The above stimuli correspond to the experiments for the pronoun *his*. The stimuli for *him* are the same except that the noun immediately following the MASK was removed. The full set of stimuli vary the stereotypical gender of the nouns.

item	sentence
1	The boy told the dad that the actor would probably blame MASK for the recent injury.
2	The man told the wizard that the nephew would protect MASK if it became necessary.
3	The doctor told the lord that the uncle would teach MASK how to drive this weekend.
4	The actor told the god that the father would buy MASK the tickets to the performance.
5	The prince told the hero that the groom might introduce MASK to the French count.
6	The lord told the master that the priest might introduce MASK to the beautiful movie star.
7	The king told the policeman that the driver would not forgive MASK for last week's disaster.
8	The waiter told the priest that the professor would take care of MASK during the holidays.
9	The actor told the hunter that the brother would protect MASK from getting hurt.
10	The nephew told the emperor that the son would entertain MASK after dinner tonight.
11	The uncle told the lord that the husband would remind MASK of the job that needed to be done.
12	The father warned the bachelor that the businessman would blame MASK for the high cost of tests.
13	The groom convinced the wizard that the boyfriend should give MASK a raise in pay.
14	The priest told the sorcerer that the congressman would probably not protect MASK under the circumstances.

item	sentence
15	The driver told the chairman that the council man would introduce MASK to the famous diplomat.
16	The professor warned the boy that the grandson would be angry with MASK for forgetting about the show.
17	The brother told the boy that the wizard might treat MASK to an expensive dinner in a nice restaurant.
18	The son told the man that the lord would introduce MASK to the band leader.
19	The husband told the doctor that the god would get MASK some lunch after the event.
20	The businessman told the actor that the hero would buy MASK twenty yards ot fine silk.
21	The boyfriend told the prince that the policeman would supply MASK with the stolen equipment.
22	The grandfather told the lord that the hunter would be proud of MASK for saving the child's life.
23	The council man told the king that the lord would be upset with MASK when the news became known.
24	The grandson told the waiter that the chairman would protect MASK if there were an investigation.

Table C.2: Templates for the stimuli for Section 5.5 and adpated from Nicol (1988). The MASK was replaced by the model specific MASK token or used the the truncation point. The full set of stimuli vary the stereotypical gender of the nouns.

item	sentence
1	When he was off work, the MASK pestered the waitress all the time.
2	When he arrived, the MASK recognized the woman at once.
3	When he was fed up, the MASK visited the girl very often.
4	When he was talking, the MASK noticed the girl at the end of the street.
5	When he was nearby, the MASK saw the lady in the park.
6	When he was lost, the MASK spotted the maid in the forest.
7	When he was bad-tempered, the MASK ignored the lady all day.
8	When he was jealous, the MASK angered the secretary more than ever.
9	When he was in residence, the MASK annoyed the princess very much.
10	When he was introduced, the MASK shook the widow by the hand.
11	When he arrived, the MASK greeted the maid kindly.
12	When he felt sad, the MASK hugged the woman gently.
13	When he was present, the MASK embarrassed the actress all the time.
14	When he was depressed, the MASK invited the lady for a drink.
15	When he was around, the MASK helped the maid all the time.
16	When he was near, the MASK approached the waitress on the plane.
17	When he was in court, the MASK trusted the lady most of all.
18	When he was poorly, the MASK depressed the actress very often.
19	When he arrived, the MASK upset the lady with the story.
20	When he was appointed, the MASK bribed the secretary shortly afterwards.
21	When he was fired, the MASK blamed the woman for the mess.
22	When he was close, the MASK recognized the girl on the path.
23	When he was retired, the MASK visited the lady almost every day.
24	When he was abroad, the MASK forgave the lady for all the troubles.
25	When he was discovered, the MASK blamed the mistress straight away.
26	When he was busy, the MASK avoided the lady as much as possible.
27	When he was twenty-one, the MASK married the bride in the cathedral.
28	When he was annoyed, the MASK disliked the maid very much.
29	When he was in church, the MASK congratulated the sister about the charity work.
30	When he was banished, the MASK missed the witch a great deal.

item	sentence
31	When he was angry, the MASK ignored the lady all the time.
32	When he was distraught, the MASK visited the nun straight away.

Table C.3: Templates for the stimuli for Section 5.6 and adpated from van Gompel and Liversedge (2003). The MASK was replaced by the model specific MASK token or used the the truncation point. The full set of stimuli vary the gender of the cataphoric pronoun.

item	cond	sentence
1	No-Gen	Before offering his son a fancy pastry, the MASK politely asked Tyler whether he had any preference.
1	B	Before offering him a fancy pastry, the MASK politely asked Tyler whether he had any preference.
1	No-Fin	Before anyone offered him a fancy pastry, the MASK politely asked Tyler whether he had any preference.
2	No-Gen	While driving his daughter to school on Friday, the MASK casually told Juan that he would pick up everyone early for a surprise.
2	B	While driving him to school on Friday, the MASK casually told Juan that he would pick up everyone early for a surprise.
2	No-Fin	While a parent drove him to school on Friday, the MASK casually told Juan that he would pick up everyone early for a surprise.
3	No-Gen	While baking his friends some cookies, the MASK happily informed Luke about all the positive new changes in his life.
3	B	While baking him some cookies, the MASK happily informed Luke about all the positive new changes in his life.
3	No-Fin	While an employee baked him some cookies, the MASK happily informed Luke about all the positive new changes in his life.
4	No-Gen	After lifting his neighbor onto the bus, the MASK carefully sat Derek down in one of the front-most seats.

item	cond	sentence
4	B	After lifting him onto the bus, the MASK carefully sat Derek down in one of the front-most seats.
4	No-Fin	After an orderly lifted him onto the bus, the MASK carefully sat Derek down in one of the front-most seats.
5	No-Gen	Before interrogating his informant about the crime, the MASK kindly offered Corey some food and a cigarette.
5	B	Before interrogating him about the crime, the MASK kindly offered Corey some food and a cigarette.
5	No-Fin	Before an officer interrogated him about the crime, the MASK kindly offered Corey some food and a cigarette.
6	No-Gen	After embarrassing his guests during the party, the MASK promptly apologized to Jeffrey for the unforgivable behavior.
6	B	After embarrassing him during the party, the MASK promptly apologized to Jeffrey for the unforgivable behavior.
6	No-Fin	After a guest embarrassed him during the party, the MASK promptly apologized to Jeffrey for the unforgivable behavior.
7	No-Gen	While helping his secretary shred the documents, the MASK tearfully admitted to Gavin that upper-level managers had been embezzling money for years.
7	B	While helping him shred the documents, the MASK tearfully admitted to Gavin that upper-level managers had been embezzling money for years.
7	No-Fin	While an intern helped him shred the documents, the MASK tearfully admitted to Gavin that upper-level managers had been embezzling money for years.
8	No-Gen	While reading his grand kids a bedtime story, the MASK gently gestured to Luis to turn off the lights.
8	B	While reading him a bedtime story, the MASK gently gestured to Luis to turn off the lights.
8	No-Fin	While someone read him a bedtime story, the MASK gently gestured to Luis to turn off the lights.
9	No-Gen	After cornering his intern next to the water cooler, the MASK loudly insulted Jorge much to everyone's horror.

item	cond	sentence
9	B	After cornering him next to the water cooler, the MASK loudly insulted Jorge much to everyone's horror.
9	No-Fin	After a colleague cornered him next to the water cooler, the MASK loudly insulted Jorge much to everyone's horror.
10	No-Gen	After weighing his patient on the scale, the MASK calmly informed Steven that he should be concerned about the onset of gout.
10	B	After weighing him on the scale, the MASK calmly informed Steven that he should be concerned about the onset of gout.
10	No-Fin	After a nurse weighed him on the scale, the MASK calmly informed Steven that he should be concerned about the onset of gout.
11	No-Gen	Before spotting his instructor at yoga class, the MASK secretly followed Christian around town.
11	B	Before spotting him at yoga class, the MASK secretly followed Christian around town.
11	No-Fin	Before anyone spotted him at yoga class, the MASK secretly followed Christian around town.
12	No-Gen	Before visiting his family on Sunday afternoons, the MASK usually called Jason up to confirm that they were still free.
12	B	Before visiting him on Sunday afternoons, the MASK usually called Jason up to confirm that he was still free.
12	No-Fin	Before friends visited him on Sunday afternoons, the MASK usually called Jason up to confirm that they were still free.

Table C.4: Templates for the stimuli for Section 5.7 and adpated from Kush and Dillon (2021). The MASK was replaced by the model specific MASK token or used the the truncation point. Experiment 1 corresponds to the No-Gen and B conditions. Experiment 2 corresponds to the No-Fin and B conditions. The full set of stimuli vary the gender of the cataphoric pronoun.

REFERENCES

Gerry Altmann and Mark Steedman. 1988. Interaction with context during human sentence processing. *Cognition*, 30(3):191–238.

Gerry T. M. Altmann. 2013. Anticipating the garden path: The horse raced past the barn ate the cake. In Montserrat Sanz, Itziar Laka, and Michael K. Tanenhaus, editors, *Language Down the Garden Path*, pages 111–130. Oxford University Press.

Aixiu An, Peng Qian, Ethan Wilcox, and Roger Levy. 2019. Representation of Constituents in Neural Language Models: Coordination Phrase as a Case Study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2888–2899, Hong Kong, China. Association for Computational Linguistics.

Suhas Arehalli and Tal Linzen. 2020. Neural Language Models Capture Some, But Not All, Agreement Attraction Effects. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, pages 370–376.

William Badecker and Kathleen Straub. 2002. The processing role of structural constraints on interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(4):748–769.

Marco Baroni. 2022. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *arXiv preprint arXiv:2106.08694*.

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence Classification with Human Attention. In *Proceedings of the*

22nd Conference on Computational Natural Language Learning, pages 302–312, Brussels, Belgium. Association for Computational Linguistics.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do Neural Machine Translation Models Learn about Morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Emily M. Bender. 2009. Linguistically Naïve != Language Independent: Why NLP Needs Linguistic Typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Jean-Phillipe Bernardy and Shalom Lappin. 2017. Using Deep Neural Networks

to Learn Syntactic Agreement. In *Linguistic Issues in Language Technology, Volume 15, 2017*. CSLI Publications.

Thomas Bever. 1970. The Cognitive Basis for Linguistic Structures. In John R. Hayes, editor, *Cognition and the Development of Language*, pages 279–352. Wiley.

Debasmita Bhattacharya and Marten van Schijndel. 2020. Filler-gaps that neural networks fail to generalize. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 486–495, Online. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8718–8735, Online. Association for Computational Linguistics.

Paul Boersma and Bruce Hayes. 2001. Empirical Tests of the Gradual Learning Algorithm. *Linguistic Inquiry*, 32(1):45–86.

James Bogen and James Woodward. 1988. Saving the Phenomena. *The Philosophical Review*, 97(3):303.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori

Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258 [cs]*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R. Bowman and George Dahl. 2021. What Will it Take to Fix Benchmarking in Natural Language Understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguis-*

tics: Human Language Technologies, pages 4843–4855, Online. Association for Computational Linguistics.

Marc Brysbaert and Don C. Mitchell. 1996. Modifier Attachment in Sentence Parsing: Evidence from Dutch. *The Quarterly Journal of Experimental Psychology Section A*, 49(3):664–695.

Joan Bybee. 2006. From Usage to Grammar: The Mind's Response to Repetition. *Language*, 82(4):711–733.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In *Proceedings of the 30th USENIX Security Symposium*, pages 2633–2650.

Manuel Carreiras and Charles Clifton. 1993. Relative Clause Interpretation Preferences in Spanish and English. *Language and Speech*, 36(4):353–372.

Manuel Carreiras and Charles Clifton. 1999. Another word on parsing relative clauses: Eyetracking evidence from Spanish and English. *Memory & Cognition*, 27(5):826–833.

Pengxiang Cheng and Katrin Erk. 2020. Attending to Entities for Better Text Understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7554–7561.

Noam Chomsky. 1959. A Review of Skinner's Verbal Behavior. *Language*, (1):26–58.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. M.I.T. Press, Cambridge.

Noam Chomsky. 1980. *Rules and Representations*. Columbia University Press, New York.

Noam Chomsky. 1981. *Lectures on Government and Binding: The Pisa Lectures*. De Gruyter Mouton.

Noam Chomsky. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Praeger, New York.

Noam Chomsky. 1995. *The Minimalist Program*. The MIT Press, Cambridge, Mass.

Noam Chomsky. 2000. Minimalist inquiries : The framework. In Rogers Martin, David Michaels, Juan Uriagereka, and Samuel Jay Keser, editors, *Step by Step: Essays on Minimalist Syntax in Honor of Howard Lasnik*. MIT Press, Cambridge, Mass.

Wing-Yee Chow, Shevaun Lewis, and Colin Phillips. 2014. Immediate sensitivity to structural constraints in pronoun resolution. *Frontiers in Psychology*, 5:630.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN Simulations of Grammaticality Judgments on Long-distance Dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shammur Absar Chowdhury and Roberto Zamparelli. 2019. An LSTM Adaptation Study of (Un)grammaticality. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 204–212, Florence, Italy. Association for Computational Linguistics.

Morten H. Christiansen. 2022. *The Language Game: How Improvisation Created Language and Changed the World*, first edition. edition. Basic Books, New York.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Charles Jr Clifton, Shelia M. Kennison, and Jason E. Albrecht. 1997. Reading the Words *Her, His, Him*: Implications for Parsing Principles Based on Frequency and on Structure. *Journal of Memory and Language*, 36(2):276–292.

Jacob Collard. 2018. Finite State Reasoning for Presupposition Satisfaction. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 53–62, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Roberta Corrigan. 1988. Who dun it? The influence of actor-patient animacy and type of verb in the making of causal attributions. *Journal of Memory and Language*, 27(4):447–465.

Roberta Corrigan. 2001. Implicit Causality in Language: Event Participants and their Interactions. *Journal of Language and Social Psychology*, 20(3):285–320.

Fernando Cuetos and Don C. Mitchell. 1988. Cross-linguistic differences in parsing: Restrictions on the use of the Late Closure strategy in Spanish. *Cognition*, 30(1):73–105.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing.

In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Jillian K. Da Costa and Rui P. Chaves. 2020. Assessing the ability of Transformer-based Neural Models to represent structurally unbounded dependencies. In *Proceedings of the Society for Computation in Linguistics*, pages 12–21, New York, NY. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Forrest Davis and Marten van Schijndel. 2020a. Discourse structure interacts with reference but not syntax in neural language models. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 396–407, Online. Association for Computational Linguistics.

Forrest Davis and Marten van Schijndel. 2020b. Interaction with Context During Recurrent Neural Network Sentence Processing. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, pages 2744–2750.

Forrest Davis and Marten van Schijndel. 2020c. Recurrent Neural Network Language Models Always Learn English-Like Relative Clause Attachment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1979–1990, Online. Association for Computational Linguistics.

Forrest Davis and Marten van Schijndel. 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, pages 1159–1171, Online. Association for Computational Linguistics.

Rodolfo Delmonte, Antonella Bristot, and Sara Tonelli. 2007. VIT – Venice Italian Treebank: Syntactic and Quantitative Features. In *Sixth International Workshop on Treebanks and Linguistic Theories*, pages 43–54.

Leon Derczynski, Hannah Rose Kirk, Abeba Birhane, and Bertie Vidgen. 2022. Handling and Presenting Harmful Text. *arXiv preprint arXiv:2204.14256 [cs].*

Timothy Desmet, Constantijn De Baecke, Denis Drieghe, Marc Brysbaert, and Wietske Vonk. 2006. Relative clause attachment in Dutch: On-line comprehension corresponds to corpus frequencies when lexical variables are taken into account. *Language and Cognitive Processes*, 21(4):453–485.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sekou Diao. 2021. mlconjug3. *Github.*

Brian Dillon, Alan Mishler, Shayne Sloggett, and Colin Phillips. 2013. Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2):85–103.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent Neural Network Grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 199–209, San Diego, California. Association for Computational Linguistics.

Chris Dyer, Gábor Melis, and Phil Blunsom. 2019. A Critical Analysis of Biased Parsers in Unsupervised Parsing. *arXiv preprint arXiv:1909.09428 [cs]*.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Jeffrey L. Elman. 1991. Distributed Representations, Simple Recurrent Networks, And Grammatical Structure. *Machine Learning*, 7(2):195–225.

Émile Enguehard, Yoav Goldberg, and Tal Linzen. 2017. Exploring the Syntactic Abilities of RNNs with Multi-task Learning. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, pages 3–14, Vancouver, Canada. Association for Computational Linguistics.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Eva M. Fernández. 2003. *Bilingual Sentence Processing: Relative Clause Attachment in English and Spanish*. John Benjamins Publishing Company.

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

Jerry A Fodor. 1983. *The Modularity of Mind*. MIT Press, Cambridge, Mass.

Stefan L. Frank and John Hoeks. 2019. The interaction between structure and meaning in sentence comprehension: Recurrent neural networks and reading times. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 377–343.

Lyn Frazier and Charles Jr Clifton. 1996. *Construal*. MIT Press.

Lyn Frazier and Janet Dean Fodor. 1978. The sausage machine: A new two-stage parsing model. *Cognition*, 6(4):291–325.

Anne Therese Frederiksen and Rachel I. Mayberry. 2021. Implicit causality biases and thematic roles in American Sign Language. *Behavior Research Methods*, 53(5):2172–2190.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2018a. The Natural Stories Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan. European Language Resources Association (ELRA).

Richard Futrell and Roger P. Levy. 2019. Do RNNs learn human-like abstract word order preferences? In *Proceedings of the Society for Computation in Linguistics*, pages 50–59.

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018b. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329 [cs]*.

Catherine Garvey and Alfonso Caramazza. 1974. Implicit Causality in Verbs. *Linguistic Inquiry*, 5(3):459–464.

Catherine Garvey, Alfonso Caramazza, and Jack Yates. 1974. Factors influencing assignment of pronoun antecedents. *Cognition*, 3(3):227–243.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Elizabeth Gilboy, Josep-Maria Sopena, Charles Clifton, and Lyn Frazier. 1995. Argument structure and association preferences in Spanish and English complex NPs. *Cognition*, 54(2):131–167.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

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

Adele E Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago.

Adele E Goldberg. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5):219–224.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren

Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Fanda Lora, Adeen Flinker, Sasha Devore, Werner Doyle, Daniel Friedman, Patricia Dugan, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. 2020. Thinking ahead: Prediction in context as a keystone of language in humans and machines. *bioRxiv*.

Olivia Guest and Andrea E. Martin. 2021. On logical inference over brains, behaviour, and artificial neural networks. *PsyArXiv*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless Green Recurrent Networks Dream Hierarchically. 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 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Joshua K. Hartshorne. 2014. What is implicit causality? *Language, Cognition and Neuroscience*, 29(7):804–824.

Joshua K. Hartshorne and Jesse Snedeker. 2013. Verb argument structure predicts implicit causality: The advantages of finer-grained semantics. *Language and Cognitive Processes*, 28(10):1474–1508.

Joshua K. Hartshorne, Yasutada Sudo, and Miki Uruwashi. 2013. Are implicit causality pronoun resolution biases consistent across languages and cultures? *Experimental Psychology*, 60(3):179–196.

Micha Heilbron, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P. de Lange. 2020. A hierarchy of linguistic predictions during natural language comprehension. *bioRxiv*.

John Hewitt and Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Dirk Hovy and Diyi Yang. 2021. The Importance of Modeling Social Factors of Language: Theory and Practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Jennifer Hu, Sherry Yong Chen, and Roger Levy. 2020a. A closer look at the performance of neural language models on reflexive anaphor licensing. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 323–333, New York, New York. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020b. A Systematic Assessment of Syntactic Generalization in Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

C.-T. James Huang. 1984. On the Distribution and Reference of Empty Pronouns. *Linguistic Inquiry*, 15(4):531–574.

Dieuwke Hupkes and Willem Zuidema. 2018. Visualisation and 'Diagnostic Classifiers' Reveal how Recurrent and Recursive Neural Networks Process Hierarchical Structure (Extended Abstract). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 5617–5621, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 310–323, New York, NY, USA. Association for Computing Machinery.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do Language Models Understand Anything? On the Ability of LSTMs to Understand Negative Polarity Items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics.

Jaap Jumelet, Willem Zuidema, and Dieuwke Hupkes. 2019. Analysing Neural Language Models: Contextual Decomposition Reveals Default Reasoning in Number and Gender Assignment. In *Proceedings of the 23rd Conference on*

Computational Natural Language Learning, pages 1–11, Hong Kong, China. Association for Computational Linguistics.

Andrew Kehler, Laura Kertz, Hannah Rohde, and Jeffrey L. Elman. 2007. Coherence and Coreference Revisited. *Journal of Semantics*, 25(1):1–44.

Andrew Kehler and Hannah Rohde. 2015. Pronominal Reference and Pragmatic Enrichment: A Bayesian Account. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, pages 1063–1068.

Andrew Kehler and Hannah Rohde. 2019. Prominence and coherence in a Bayesian theory of pronoun interpretation. *Journal of Pragmatics*, 154:63–78.

Yova Kementchedjhieva, Mark Anderson, and Anders Søgaard. 2021. John praised Mary because *he*? Implicit Causality Bias and Its Interaction with Explicit Cues in LMs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4859–4871, Online. Association for Computational Linguistics.

Shelia M. Kennison. 2003. Comprehending the pronouns *her*, *him*, and *his*: Implications for theories of referential processing. *Journal of Memory and Language*, 49(3):335–352.

Yoon Kim, Alexander Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and Gábor Melis. 2019. Unsupervised Recurrent Neural Network Grammars. 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 1105–1117, Minneapolis, Minnesota. Association for Computational Linguistics.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533, San Diego, California. Association for Computational Linguistics.

Jordan Kodner and Nitish Gupta. 2020. Overestimation of Syntactic Representation in Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1757–1762, Online. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Dave Kush and Brian Dillon. 2021. Principle B constrains the processing of cataphora: Evidence for syntactic and discourse predictions. *Journal of Memory and Language*, 120:104254.

Dave Kush, Jeffrey Lidz, and Colin Phillips. 2015. Relation-sensitive retrieval: Evidence from bound variable pronouns. *Journal of Memory and Language*, 82:18–40.

Dave Kush and Colin Phillips. 2014. Local anaphor licensing in an SOV language: Implications for retrieval strategies. *Frontiers in Psychology*, 5:1252.

Dave W. Kush. 2013. *Respecting Relations: Memory Access and Antecedent Retrieval in Incremental Sentence Processing*. Ph.D. thesis, University of Maryland.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H.B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13):1–26.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41(5):1202–1241.

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

David K. Lewis. 1975. Languages and Language. In Keith Gunderson, editor, *Minnesota Studies in the Philosophy of Science*, volume Volume VII, pages 3–35. University of Minnesota Press, Minneapolis.

Richard L. Lewis and Shravan Vasishth. 2005. An Activation-Based Model of Sentence Processing as Skilled Memory Retrieval. *Cognitive Science*, 29(3):375–419.

Bingzhi Li and Guillaume Wisniewski. 2021. Are Neural Networks Extracting Linguistic Properties or Memorizing Training Data? An Observation with a Multilingual Probe for Predicting Tense. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3080–3089, Online. Association for Computational Linguistics.

Tal Linzen and Marco Baroni. 2021. Syntactic Structure from Deep Learning. *Annual Review of Linguistics*, 7(1):195–212.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4(0):521–535.

Tal Linzen and Brian Leonard. 2018. Distinct patterns of syntactic agreement errors in recurrent networks and humans. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 6790–6795.

Chi-Ming Louis Liu. 2014. *A Modular Theory of Radical Pro Drop*. Ph.D. thesis, Harvard University.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692 [cs]*.

Peter Ludlow. 2011. *The Philosophy of Generative Linguistics*. Oxford University Press, Oxford.

Maryellen C. MacDonald. 1994. Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, 9(2):157–201.

Brian MacWhinney, Elizabeth Bates, and Reinhold Kliegl. 1984. Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior*, 23(2):127–150.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets Construction Grammar. In *Proceedings of*

the 28th International Conference on Computational Linguistics, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lucia Mannetti and Eraldo De Grada. 1991. Interpersonal verbs: Implicit causality of action verbs and contextual factors: Implicit causality of action verbs. *European Journal of Social Psychology*, 21(5):429–443.

Rebecca Marvin and Tal Linzen. 2018. Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Laia Mayol. 2012. An account of the variation in the rates of overt subject pronouns in Romance. *Spanish in Context*, 9(3):420–442.

J. L. McClelland, Mark St John, and Roman Taraban. 1989. Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes*, 4(3-4):SI287–SI335.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2018. Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 2096–2101.

R. Thomas McCoy, Erin Grant, Paul Smolensky, Thomas L. Griffiths, and Tal Linzen. 2020. Universal linguistic inductive biases via meta-learning. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, pages 737–743.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2021. How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN. *arXiv preprint arXiv 2111.09509*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. WikiText Dataset. Technical report.

Sabrina J. Mielke. 2016. Language diversity in ACL 2014-2016.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635, Online. Association for Computational Linguistics.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-Linguistic Syntactic Evaluation of Word Prediction Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First Align, then Predict: Understanding the Cross-Lingual Ability of Multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for*

Computational Linguistics: Main Volume, pages 2214–2231, Online. Association for Computational Linguistics.

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021. Refining Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3710–3723, Online. Association for Computational Linguistics.

Binh Ngo and Elsi Kaiser. 2020. Implicit Causality: A Comparison of English and Vietnamese Verbs. In *Proceedings of the 43rd Annual Penn Linguistics Conference*, pages 179–186.

Janet L. Nicol. 1988. *Coreference Processing during Sentence Comprehension*. Thesis, Massachusetts Institute of Technology.

Janet L. Nicol and David Swinney. 1989. The role of structure in coreference assignment during sentence comprehension. *Journal of Psycholinguistic Research*, 18(1):5–19.

Janet L. Nicol and David A. Swinney. 2003. The Psycholinguistics of Anaphora. In Andrew Barss, editor, *Anaphora*, pages 72–104. Blackwell Publishing Ltd, Malden, MA, USA.

Ricardo Otheguy, Ana Celia. Zentella, and David. Livert. 2008. Language and Dialect Contact in Spanish in New York: Toward the Formation of a Speech Community. *Language*, 83(4):770–802.

Dario Paape and Shravan Vasishth. 2022. Estimating the true cost of garden-pathing: A computational model of latent cognitive processes. *PsyArXiv*.

Ludovica Pannitto and Aurélie Herbelot. 2020. Recurrent babbling: Evaluating the acquisition of grammar from limited input data. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 165–176, Online. Association for Computational Linguistics.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. UmBERTo: An Italian Language Model trained with Whole Word Masking.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 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.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using Priming to Uncover the Organization of Syntactic Representations in Neural Language Models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Alan Prince and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell Pub., Malden, MA.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation.

In *Proceedings of the 38th International Conference on Machine Learning*, pages 8821–8831. PMLR.

Giulio Ravasio and Leonardo Di Perna. 2020. GilBERTo: An Italian pretrained language model based on RoBERTa.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. 2019. Studying the Inductive Biases of RNNs with Synthetic Variations of Natural Languages. 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 3532–3542, Minneapolis, Minnesota. Association for Computational Linguistics.

Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 2018. Can LSTM Learn to Capture Agreement? The Case of Basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107, Brussels, Belgium. Association for Computational Linguistics.

Chiara Reali, Yulia Esaulova, Anton Öttl, and Lisa von Stockhausen. 2015. Role descriptions induce gender mismatch effects in eye movements during reading. *Frontiers in Psychology*, 6.

Tanya Reinhart and Eric Reuland. 1993. Reflexivity. *Linguistic Inquiry*, 24(4):657–720.

Georges Rey. 2020. *Representation of Language: Philosophical Issues in a Chomskyan Linguistics*. Oxford University Press, Oxford, New York.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Luigi Rizzi. 1986. Null Objects in Italian and the Theory of pro. *Linguistic Inquiry*, 17(3):501–557.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Hannah Rohde, Roger Levy, and Andrew Kehler. 2011. Anticipating explanations in relative clause processing. *Cognition*, 118(3):339–358.

Manuel Romero. 2020. RuPERTa: The Spanish RoBERTa.

Mats Rooth. 2017. Finite state intensional semantics. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long Papers*.

John Robert Ross. 1967. *Constraints on Variables in Syntax*. Ph.D. thesis, MIT, Cambridge, Massachusetts.

David E. Rumelhart and J. L. McClelland. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, Mass.

Soo Hyun Ryu and Richard Lewis. 2021. Accounting for Agreement Phenomena in Sentence Comprehension with Transformer Language Models: Effects of Similarity-based Interference on Surprisal and Attention. In *Proceedings of the*

Workshop on Cognitive Modeling and Computational Linguistics, pages 61–71, Online. Association for Computational Linguistics.

Ferdinand de Saussure. 1983. *Course in General Linguistics*. Duckworth, London. Trans. Roy Harris.

Christoph Scheepers. 2003. Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition*, 89(3):179–205.

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. 2020. Artificial Neural Networks Accurately Predict Language Processing in the Brain. *bioRxiv*.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5387–5403, Online. Association for Computational Linguistics.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D Manning. 2014. A Gold Standard Dependency Corpus for English. In *Ninth International Conference on Language Resources and Evaluation*.

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

Ionut-Teodor Sorodoc, Kristina Gulordava, and Gemma Boleda. 2020. Probing for Referential Information in Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4177–4189, Online. Association for Computational Linguistics.

Patrick Sturt. 2003. The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, 48(3):542–562.

Whitney Tabor, Bruno Galantucci, and Daniel Richardson. 2004. Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, 50(4):355–370.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, Marrakech, Morocco. European Language Resources Association (ELRA).

Anne-Michelle Tessier. 2009. Frequency of violation and constraint-based phonological learning. *Lingua*, 119(1):6–38.

James Tollefson. 2002. The language debates: Preparing for the war in Yugoslavia, 1980-1991. 2002(154):65–82.

Shiva Upadhye, Leon Bergen, and Andrew Kehler. 2020. Predicting Reference: What do Language Models Learn about Discourse Models? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 977–982, Online. Association for Computational Linguistics.

Roger P.G. van Gompel and Simon P. Liversedge. 2003. The Influence of Morphological Information on Cataphoric Pronoun Assignment. *Journal of Experimental Psychology. Learning, Memory & Cognition*, 29(1):128–139.

Marten van Schijndel and Tal Linzen. 2018a. Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 692–697.

Marten van Schijndel and Tal Linzen. 2018b. A Neural Model of Adaptation in Reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4704–4710, Brussels, Belgium. Association for Computational Linguistics.

Marten van Schijndel and Tal Linzen. 2021. Single-Stage Prediction Models Do Not Explain the Magnitude of Syntactic Disambiguation Difficulty. *Cognitive Science*, 45(6):e12988.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5831–5837, Hong Kong, China. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019a. Investigating BERT's Knowledge of Language: Five Analysis Methods

with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020a. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019b. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020b. Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 217–235, Online. Association for Computational Linguistics.

Ethan Wilcox, Richard Futrell, and Roger Levy. 2021a. Using Computational Models to Test Syntactic Learnability. *LingBuzz*.

Ethan Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020a. On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, pages 1707–1713.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019a. Hierarchical Representation in Neural Language Models: Suppression and Recovery of Expectations. In

Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 181–190, Florence, Italy. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019b. What Syntactic Structures block Dependencies in RNN Language Models? In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 1199–1205.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN Language Models Learn about Filler–Gap Dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Ethan Wilcox, Peng Qian, Richard Futrell, Ryosuke Kohita, Roger Levy, and Miguel Ballesteros. 2020b. Structural Supervision Improves Few-Shot Learning and Syntactic Generalization in Neural Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4640–4652, Online. Association for Computational Linguistics.

Ethan Wilcox, Pranali Vani, and Roger Levy. 2021b. A Targeted Assessment of Incremental Processing in Neural Language Models and Humans. 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 939–952, Online. Association for Computational Linguistics.

Elyce Dominique Williams. 2020. Language Experience Predicts Pronoun Comprehension in Implicit Causality Sentences. Master's thesis, University of North Carolina at Chapel Hill.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Eunkyung Yi and Jean-Pierre Koenig. 2020. Grammar modulates discourse expectations: Evidence from causal relations in English and Korean. *Language and Cognition*, pages 1–29.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.