Incremental Processing of Principle B: Mismatches Between Neural Models and Humans

Forrest Davis
Department of Linguistics and Philosophy
Massachusetts Institute of Technology
forrestd@mit.edu

Abstract

Despite neural language models qualitatively capturing many human linguistic behaviors, recent work has demonstrated that they underestimate the true processing costs of ungrammatical structures. We extend these more fine-grained comparisons between humans and models by investigating the interaction between Principle B and coreference processing. While humans use Principle B to block certain structural positions from affecting their incremental processing, we find that GPT-based language models are influenced by ungrammatical positions. We conclude by relating the mismatch between neural models and humans to properties of training data and suggest that certain aspects of human processing behavior do not directly follow from linguistic data.

1 Introduction

Neural models trained on text data alone have been shown to qualitatively capture aspects of a large variety of human linguistic behaviors (e.g., Gulordava et al., 2018; Wilcox et al., 2019; Warstadt et al., 2020; Hu et al., 2020; Jumelet et al., 2021). Investigations have evaluated a range of levels of linguistic knowledge, including: i) syntax (Marvin and Linzen, 2018; Warstadt et al., 2019; Wilcox et al., 2019, 2021a), ii) semantics (Pannitto and Herbelot, 2020; Misra et al., 2020), and iii) discourse structure and pragmatics (Schuster et al., 2020; Davis and van Schijndel, 2020).

Recent work has placed increased attention on finer-grained comparisons between neural models and humans (e.g., van Schijndel and Linzen, 2021; Wilcox et al., 2021b; Paape and Vasishth, 2022). The growing consensus is that neural models underestimate the processing costs seen with humans, while nonetheless capturing the broad patterns (see Wilcox et al., 2021b). The present study adds to this literature by comparing the incremental processing of coreference in humans and neural models.

While coreference, more generally, is modulated by discourse, pragmatics, and information structure (e.g., Arnold, 1998, 2001; Rohde et al., 2006; Hartshorne, 2014; Rohde and Kehler, 2014), there are sentential restrictions on coreference that have immediate effects on human incremental processing (e.g., Nicol, 1988; Clifton et al., 1997; Sturt, 2003; Chow et al., 2014). This study finds that, contrary to humans, autoregressive neural models do not similarly restrict their behavior in coreference processing.

In particular, the present study investigated the interaction between the Binding Principles, articulated in Chomsky (1981), and incremental processing. Binding Principles account for the constrained distribution of pronouns (and anaphora) and their possible linguistic antecedents:

(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

Roughly, Principle A excludes examples like *John thinks that Mike hates himself* from meaning that "John thinks that Mike hates John". Conversely, Principle B excludes examples like *John thinks that Mike hates him* from meaning that "John thinks that Mike hates Mike". Finally, Principle C excludes *He hates John* from meaning "John hates John". These principles are mediated by a structural relation, c-command, rather than linear order. While the specific binding conditions have been refined within syntactic theory (e.g., Reinhart and Reuland, 1993), we focused here on the empirical results concerning Principle B and incremental processing, putting aside explicit theoretical commitments.

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

Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL), pages 144 - 156

In (2), despite *him* agreeing in gender with *Bill*, *Clark*, and *Robert*, only two of these are possible antecedents of *him*: *Bill* and *Clark*. Principle B blocks the structural location occupied by *Robert* from serving as an antecedent of *him*. In human incremental processing, this restriction has immediate effects, preventing the gender of this embedded subject from influencing the processing of the pronoun (see Chow et al., 2014). Moreover, Principle B can restrict the prediction of nouns following certain cataphoric pronouns – pronouns that occur before their coreferring noun phrase (Kush and Dillon, 2021). For example, in (3), *him* can only corefer with *Mark* and not *Michael*. In human incremental processing, the cataphoric pronoun *him* has no effect on the processing of the subject (e.g., *Michael*).

(3) Before offering him a fancy pastry, Michael politely asked Mark for help.

In what follows, we evaluate whether GPT-like autoregressive neural models use Principle B to restrict their incremental processing like humans. Specifically, we investigated two broad effects of Principle B: i) its interaction with "vanilla" pronouns (as in (2)), and ii) its interaction with cataphora (as in (3)).

While models appear to learn aspects of Principle B (treating apparent violations in unique ways), we find that neural models, in contrast to humans, do not categorically ignore structural positions blocked by Principle B. Ultimately, the present study suggests that, beyond underestimating the processing costs seen in humans, models fail, at least in some cases, to learn qualitatively similar patterns to humans. This suggests, in turn, that certain aspects of human parsing behavior are not directly evidenced in linguistic data.

2 Background

In human coreference processing, a major question is whether antecedent retrieval, triggered by the presence of a pronoun, is restricted first by agreement features (e.g., gender, number), returning possibly ungrammatical antecedents, or by structural constraints, like Principle B, which serve as an initial filter. As an illustration consider the following set of stimuli discussed in Chow et al. (2014):

(4) a. John thought that Bill liked him.
b. John thought that Mary liked him.
c. Jane thought that Bill liked him.
d. Jane thought that Mary liked him.

If Principle B immediately restricts the set of possible antecedents of *him*, then we would expect the reading times at *him* to be the same for (4-a) and (4-b), as in both cases the structurally licit antecedent agrees in gender. If instead structurally ungrammatical antecedents can influence the immediate processing of *him*, then we would expect that (4-a)–(4-c) would pattern together, to the exclusion of (4-d), where no antecedent is given in the linguistic context. Put another way, whether the structurally ungrammatical antecedent influences reading times at *him* is indicative of the status of Principle B in human linguistic processing.

The bulk of work investigating these, and similar constructions, has found that structural constraints like Principle B do immediately influence human incremental processing (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.[1]

Within work in natural language processing, existing models have been claimed to capture aspects of Principle A (e.g., Warstadt et al., 2020; Hu et al., 2020). Principle C has received less attention, though see Mitchell et al. (2019) which found that LSTM language models failed to obey Principle C. Coreference, more broadly, has also been explored, with results suggesting that models encode features of coreference resolution (e.g., Sorodoc et al., 2020) and the interaction of implicit causality and pronouns (verb biases that influence preferred antecedents for pronouns; Upadhye et al., 2020; Davis and van Schijndel, 2021; Kementchedjhieva et al., 2021).

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 coreferring with certain antecedents (and thus fully verify whether they have learned Principle B, or even Principle A), we can compare the difference in model behavior conditioned on minimally contrastive stimuli. In fact, human online sentence comprehension stud-

[1] However, some other work has suggested that grammatically illicit antecedents can in fact have measurable effects (e.g., Badecker and Straub, 2002; Kennison, 2003). Such effects may be capturing later stages of processing (see Sturt, 2003). Nevertheless, the plurality of the evidence suggests that Principle B has immediate effects on human processing.

ies are similarly limited. Since we cannot directly measure the content retrieved in reading a pronoun, online reading times are taken as a proxy for the consideration of certain antecedents.

3 Neural Models and Measures

We analyzed four autoregressive models with GPT-like architectures: GPT-2 XL (1.5B parameters; Radford et al., 2019), GPT-Neo (2.7B parameters; Black et al., 2021), GPT-J (6B parameters; Wang and Komatsuzaki, 2021), and GPT-3 (175B parameters; Brown et al., 2020). GPT-2 XL, GPT-Neo, and GPT-J were accessed via HuggingFace (Wolf et al., 2020), and GPT-3 by using OpenAI's API.[2]

In evaluating model performance, we used *surprisal* (Hale, 2001; Levy, 2008):

$$-\log \text{Prob}(\text{word}|\text{context}) \quad (1)$$

Surprisal has a linear relationship with human reading times (Smith and Levy, 2013). We follow a growing body of work in utilizing this relationship to compare the behavior of neural models and humans (e.g., van Schijndel and Linzen, 2021; Wilcox et al., 2021b).[3]

To aid the interpretation of the results, we calculated by-item gender mismatch effects (GMMEs). GMMEs are used in human experiments to index the increased cost in processing incurred when encountering a pronoun (or a postcedent, in the case of cataphoric pronoun processing) that was not expected (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. We calculated two classes of GMMEs for neural models targeting gender prediction for, i) "vanilla" pronouns, and ii) subjects after reading cataphoric pronouns.

For predictions about upcoming pronouns, consider:

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

To calculate the GMME for (5), we took the difference between the surprisal for *him* in (5-a) and the surprisal for *him* in (5-b). More generally, we calculated a GMME by taking the difference in the surprisal of the target (either a pronoun or the subject noun) between minimal pairs. 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., 2014).[4]

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 (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 and Pronouns

Recall, humans restrict their incremental processing of coreference to just those antecedents which are grammatically licensed (e.g., Chow et al., 2014). That is, in sentences like *Fred thought Amy hated him*, *him* cannot be co-indexed with the structural position that *Amy* occupies, and thus, the gender of *Amy* does not hinder the processing of *him*. In this section, we evaluated the ability of GPT-like autoregressive neural models to replicate this qualitative effect across four experimental conditions.

4.1 Stimuli

In this section, we consider four experiments:

(7) **Experiments**

[2]We used the version of GPT-3 called text-davinci-002. All the stimuli, results, and scripts for recreating the statistics and figures can be found at https://github.com/forrestdavis/PrincipleB.

[3]For a more explicit comparison between human self-paced reading times and neural models see Section 6.1.

[4]Because the feminine pronoun *her* is ambiguous between a possessive and an object pronoun when processing left to right (e.g., *Sue loves her* and *Sue loves her friend*) only masculine pronouns were evaluated in pronoun prediction.

[5]All subject nouns investigated were encoded by the neural models as single tokens rather than being split into multiple tokens as in Randolf mapping to 'Rand' + 'olf' in GPT-J.

a. SIMPLE SUBJECT: Single clause with simple subject
b. COMPLEX SUBJECT: Single clause with complex subject containing a prepositional phrase
c. 2NP: Clause with embedding and simple subjects
d. 3NP: Clause with embedding and simple subjects and an object

Examples of each are included below:

(8) **Stimuli Examples**

a. SIMPLE SUBJECT: The boy meets him.
b. COMPLEX SUBJECT: The story about Eric hurt him.
c. 2NP: Jason hadn't expected that Adam was investigating him.
d. 3NP: Liam advised the nephew that Patrick can praise him.

We used the data generation scripts and vocabulary provided with the BLiMP dataset to create our stimuli (Warstadt et al., 2020). The sentences are all grammatical and generally semantically felicitous (despite certain interpretations being blocked by Principle B). The stimuli for COMPLEX SUBJECT always had a subject comprised of "the X about…", where X ranged over inanimate nouns like *book* or *story*.[6]

There were 1000 base sentences for each experiment, with each sentence having exponents that varied gender in all relevant positions (e.g., (8-c) has four forms varying whether the matrix subject is *Jason* or *Amanda* and whether the embedded subject is *Adam* or *Victoria*).[7] The applicability of Principle B varied by experiment. For SIMPLE SUBJECT, Principle B blocks co-indexation between the subject and the object pronoun. For COMPLEX SUBJECT, Principle B does not block co-indexation between the lower noun (e.g., *Eric* in (8-b)) and the pronoun. Principle B, however, does block the higher nouns (e.g., *the story* in (8-b)) from co-indexing with the pronoun *him*.[8] For 2NP and 3NP, Principle B blocks co-indexation between the

[6]The full set contained *book, pamphlet, brochure, play, movie, newspaper article, story, essay, report, documentary, commentary,* and *show*.

[7]No noun was repeated in a single sentence. That is, there were no sentences like *The man advised the nephew that the man can praise him*.

[8]Additionally, all higher nouns were inanimate, again blocking the applicability of *him*.

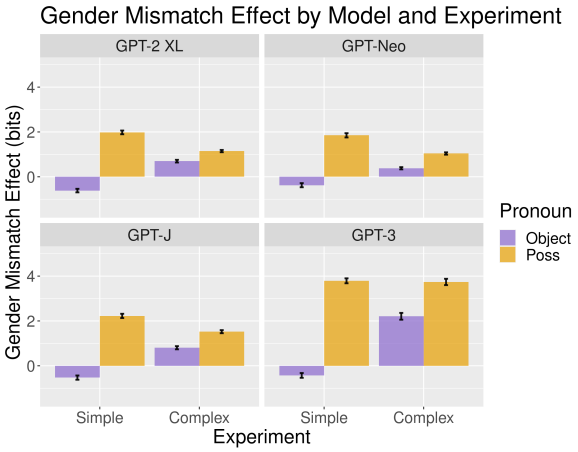

Figure 1: GMME for object pronoun (*him*) and possessive pronoun (*his*) for each neural model by two conditions: i) SIMPLE SUBJECT, and ii) COMPLEX SUBJECT (e.g., (*Bill*|*The book about Bill*) *worried him*). Error bars are 95% confidence intervals.

embedded subject (e.g., *Adam* in (8-c) and *Patrick* in (8-d)) and the pronoun, but not the matrix subject (e.g., *Jason* in (8-c) and *Liam* in (8-d)) or matrix object for 3NP (e.g., *the nephew* in (8-d)). If neural models patterned like humans, then we should find no GMME when Principle B blocks co-indexation, and positive GMMEs elsewhere.

4.2 Simple Sentences and Pronoun Prediction

First, we investigated the influence on pronoun prediction that subjects had in single clause constructions (the SIMPLE SUBJECT and COMPLEX SUBJECT experiments; see (8-a) and (8-b) above for the relevant contrasts).

Results grouped by model, condition, and pronoun are given in Figure 1. Statistical analyses were conducted via linear-mixed effects models.[9]

Starting with the results for possessive pronouns, we found that all models showed a positive GMME. That is, models expected possessive pronouns to agree in gender with the subject, both in simple sentences (e.g., *Fred worried his…*) and sentences with complex subjects (e.g., *The book about Fred worried his…*).

For object pronouns, GMME differed by subject type. For complex subjects, where co-indexation between the object pronoun and the lower noun (e.g., *Fred* in *The book about Fred*) is possible,

[9]We used lmer (version 1.1.30; Bates et al., 2015) and lmerTest (version 3.1.3; Kuznetsova et al., 2017) in R. Models were fit to predict the surprisal of the pronoun *him* or *his* with a main effect of condition (i.e. whether the noun matched the gender of the pronoun) with by-item random intercepts.

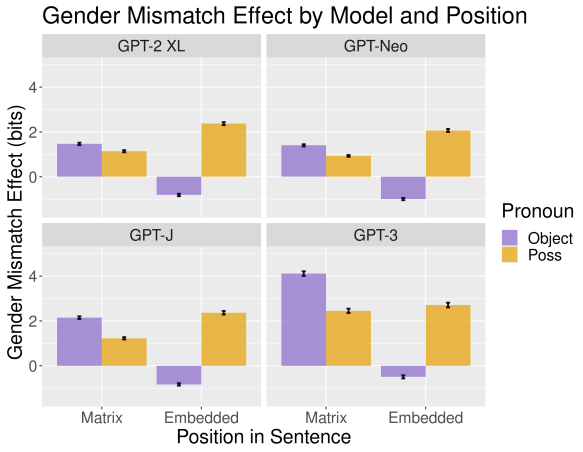

Figure 2: GMME for object pronoun (*him*) and possessive pronoun (*his*) for each neural model by whether i) the matrix subject, or ii) the embedded subject agrees in gender (e.g., (*Bill*|*Hannah*) *thinks that* (*Mark*|*Sue*) *hates him*). Error bars are 95% confidence intervals.

models again exhibited a positive GMME, suggesting that agreement between the object pronoun and the lower noun was expected. For simple subjects, where co-indexation between the subject and the pronoun is **not** possible (e.g., *him* cannot refer to *Fred* in *Fred worried him*), a negative GMME was obtained. That is, despite the subject not being a possible coreferent for the object pronoun, the gender of the subject (negatively) influenced the surprisal of the object pronoun.

4.3 Multiple NPs

We found evidence that, in cases where co-indexation is blocked by Principle B, models expected pronouns to mismatch with the gender of the antecedent. While suggesting that models consider antecedents that humans do not, it nonetheless suggests models capture aspects of the ungrammaticality of violations of Principle B. In this section, we evaluated models on more complex sentences containing two or three noun phrase antecedents (the 2NP and 3NP experiments; see examples (8-c) and (8-d) in Section 4.1 for the relevant contrasts).

Results for the 2NP case are given in Figure 2 (with results for the 3NP case given in Figure 7 in Appendix A). Statistical analyses were conducted via linear-mixed effects models.[10] Starting with the results for possessive pronouns, in both conditions, all models exhibited a positive GMME in all positions (e.g., matrix subject, embedded subject). That is, models predicted that possessive pronouns would agree with the antecedent nouns.

For object pronouns, we again found a mismatch in the direction of the GMME conditioned on the structural position of the relevant antecedent. When co-indexation is grammatically licensed (e.g., *him* can refer to *Bill* in *Bill knows that Mary loves him*), a positive GMME was obtained for all models. In cases where Principle B blocks co-indexation, all models exhibited a negative GMME instead. As in Section 4, this suggests that grammatically unavailable antecedents influenced the surprisal of object pronouns contrary to the results obtained in human incremental processing.

4.4 Interim Discussion

Broadly, the above experiments demonstrated that neural models exhibited GMMEs when pronouns mismatched in gender with preceding nouns. For the possessive pronoun *his*, this amounted to positive GMMEs across-the-board. That is, mismatches in gender between *his* and any antecedent increased the surprisal of *his*. For the object pronoun *him*, the GMME interacted with Principle B. Positive GMMEs were obtained when grammatically licit antecedents mismatched in gender, suggesting models predicted *him* to agree with these antecedents. However, when Principle B blocked the structural position from permitting co-indexation between the antecedent and the object pronoun, a negative GMME was obtained. That is, models expected *him* to mismatch in gender with grammatically unavailable antecedents.

As evidenced by the COMPLEX SUBJECT experiment, this negative GMME is not merely a dispreference for local agreement with object pronouns. For sentences like *The book about Fred surprised him*, the more recent noun in linear order agrees in gender with *him*, but we found a positive GMME. Rather, neural models appear to have learned, at least some, aspects of Principle B (in so far as certain structural positions are marked). However, the negative GMME was unexpected given the findings in the literature surrounding incremental processing of such constructions in English. Ultimately, neural models appear to use information in prediction that the human parser does not.

[10] Models were fit to predict the surprisal of the pronoun *him* or *his* with an interaction between the matrix subject gender (i.e. whether it matched with the pronoun) and the embedded subject gender, in the two noun phrase case, or the matrix subject gender, the matrix object gender, and the embedded subject gender (e.g., *Fred*|*Mary told Mark*|*Karen that Frank*|*Sue hated him*), in the three noun phrase case, and with by-item random intercepts.

5 Principle B and Cataphora

The above section explored the role Principle B plays in pronoun prediction for GPT-like neural models, finding a qualitative mismatch between the incremental processing of neural models and humans. Recent work in psycholinguistics has also demonstrated that Principle B can restrict the prediction of subjects following cataphoric object pronouns (Kush and Dillon, 2021).

(9) a. While baking him some cookies, Nicholas chatted with Mark.
b. While an employee baked him some cookies, Nicholas happily chatted with Mark.

In (9), *him* is a cataphoric pronoun – the noun phrase it corefers with comes later in the sentence. While *him* can be co-indexed with *Nicholas* in (9-b) (meaning Nicholas had some cookies baked for him), *him* cannot be co-indexed with *Nicholas* in (9-a).[11] Principle B excludes this latter co-indexation.[12] Kush and Dillon (2021) found that humans exhibited a GMME at the subject (e.g., *Nicholas*) only in cases where co-indexation between the catphoric *him* and the subject was possible (e.g., (9-b)). As with "vanilla" pronouns, it seems, then, that Principle B immediately restricts the human parser, such that grammatically unavailable structural positions are ignored.

In the following section, we evaluated whether neural models patterned like humans in this respect. That is, whether models exhibited a GMME only in cases where Principle B did not block co-indexation. First, we also verified that the neural models could use cataphoric pronouns to restrict the prediction of subjects more generally.

5.1 Stimuli

In this section, we consider two experiments:

(10) **Experiments**
a. SUBJECT CATAPHORA: Sentences with a cataphoric subject pronoun
b. OBJECT CATAPHORA: Sentences with a cataphoric object pronoun

[11] A natural interpretation of (9-a) is that Nicholas was baking cookies for Mark while chatting with Mark

[12] Obligatory control of the PRO in the adjunct is also implicated by this construction. We abstract from the relevant syntactic analysis here, and instead focus on the empirical findings from human experiments (for full discussion see Kush and Dillon, 2021, and references therein).

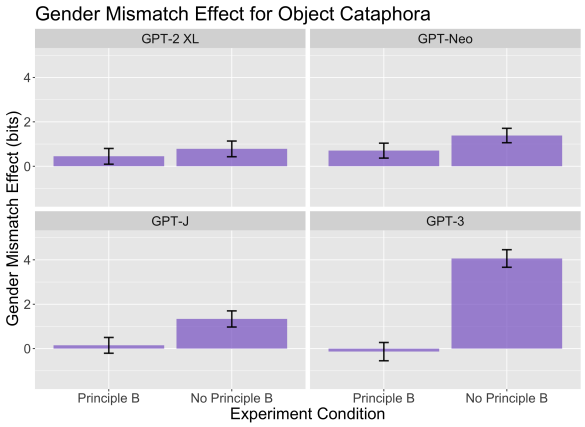

Figure 3: GMME for subject following a cataphoric object pronoun (*him*) for each neural model by whether Principle B applies (e.g., (*While driving him*|*While someone drove him*), (*Bill*|*Sue*)). Stimuli adapted from Kush and Dillon (2021). Error bars are 95% confidence intervals.

Examples of each are included below.

(11) **Stimuli Examples**
a. SUBJECT CATAPHORA: When **he** was off work, Richard...
b. OBJECT CATAPHORA: While driving **him** to school on Friday, Thomas...

For SUBJECT CATAPHORA, we used the 32 stimuli from Experiment 1 in van Gompel and Liversedge (2003). The gender of the cataphoric pronoun and the matrix subject (e.g., *he* and *Richard* in (11-a)) were manipulated resulting in male and female versions of each. Moreover, for each stimulus in van Gompel and Liversedge (2003), we evaluated models on ten unique subjects per gender.

For OBJECT CATAPHORA, we drew on the 24 stimuli from Experiment 2 in Kush and Dillon (2021), which were already balanced for gender (i.e. 12 with *him*). As with SUBJECT CATAPHORA, the experiment manipulated the gender match between the cataphoric pronoun and the subject noun. Additionally, Kush and Dillon (2021) manipulated whether Principle B applied to the construction. For instance, Principle B applies in (11-b), blocking *him* from co-indexing with *Thomas*. However, a minimal different string, *While a parent drove him to School on Friday, Thomas...*, does not implicate Principle B. We again evaluated models on ten unique subject nouns per sentence.

In this section, Principle B was only relevant for *Object Cataphora*, with *Subject Cataphora* serving

as a baseline to ensure that models can, in fact, use cataphoric pronouns to predict the gender of upcoming subjects.

5.2 Simple Subject Cataphora

We turn first to the ability of neural models to modulate their predictions of upcoming subjects by the presence of cataphoric subject pronouns (see (11-a) for a relevant example). Results are given in Figure 8 of Appendix A, and statistical analyses were conducted via linear-mixed effects models.[13] All models exhibited a positive GMME, suggesting that models use cataphoric pronouns to constrain upcoming predictions about the gender of nouns.

5.3 Cataphora and Principle B

Given that neural models can use cataphoric pronouns in prediction, we evaluated whether models capture the interaction of cataphoric processing and Principle B (see Section 5.1 for discussion of the relevant contrast). Results are given by model and experimental condition in Figure 3. Statistical significance was determined via linear-mixed effects models.[14]

Recall, that humans exhibit a GMME only in the case that Principle B does not block coreference between the cataphoric pronoun and the subject (e.g., *him* cannot be co-indexed with *Fred* in *While driving him to the store, Fred...*). If neural models capture this aspect of human incremental processing, a GMME should be obtained only in cases where Principle B is not active. We found, however, that not all models captured this distinction.

GPT-3 and GPT-J demonstrated no significant GMME in cases where Principle B blocked coreference, in line with humans. GPT-2 XL and GPT-Neo, on the other hand, had a positive GMME suggesting that models used the gender of the cataphoric pronoun to predict the gender of the subject. That is, the models predicted that the gender of the subject would agree with the cataphoric pronoun, despite co-indexation being ungrammatical for humans. When Principle B was not implicated, all models showed a positive GMME suggesting that, as with subject cataphora, object cataphora can restrict the prediction of subjects.

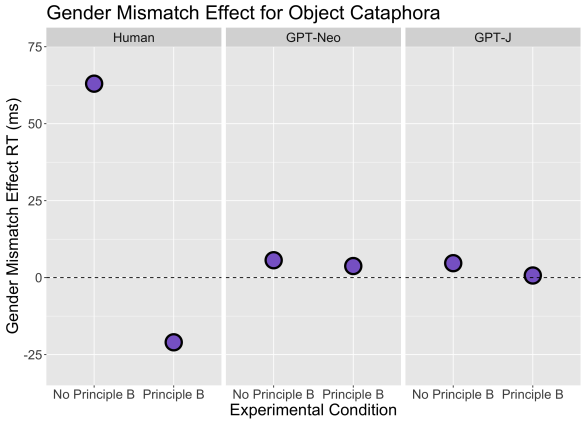

Figure 4: Mean GMME for subject following a cataphoric object pronoun (*him*) for humans (reported in Experiment 2 of Kush and Dillon (2021)), GPT-Neo, and GPT-J. Predicted reading times (in milliseconds) for the neural models were obtained by fitting the self-paced reading times for the fillers following the methodology outlined in van Schijndel and Linzen (2021).

6 General Discussion

This study investigated whether autoregressive neural models displayed similar incremental coreference processing to humans. Specifically, we examined the interaction between Principle B and coreference processing with two broad case studies: i) "vanilla" pronouns (where the antecedent precedes the pronoun), and ii) cataphoric pronouns (where the pronoun precedes its coreferring noun phrase). For the first case study, we found that the pronoun predictions of all models were influenced by structural positions deemed ungrammatical by Principle B, inconsistent with the incremental processing behavior of humans. For the second case study, we found that two of the four models (GPT-J and GPT-3), displayed human-like processing behavior in predicting subjects after cataphoric object pronouns (e.g., *him*), specifically with Principle B blocking the influence of the pronoun on the prediction of the later subject.

Three questions remain concerning the behavior of neural models: 1) how closely do models predict the observed processing cost in human studies, 2) why do GPT-J and GPT-3, and not the other models, pattern like humans in cataphoric processing, and 3) why do models consider ungrammatical antecedents in their incremental processing of pronouns.

[13] Models were fit to predict the surprisal of the subject noun with a main effect of contrast (whether the cataphoric pronoun agreed with the subject) and by-item and by-gender (*he* or *she*) random intercepts.

[14] Models were fit to predict the surprisal of the subject noun with an interaction of the gender agreement of the cataphoric pronoun (i.e. whether the pronoun and subject agreed in gender) and the presence of Principle B (i.e. whether co-indexation was possible between the cataphoric pronoun and the subject) with by-item random intercepts.

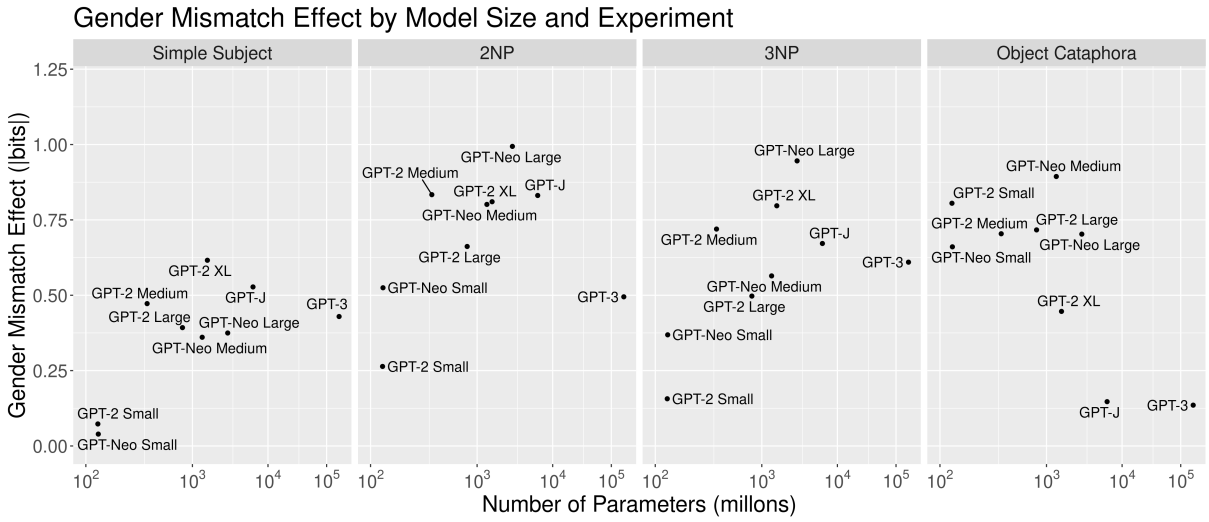

Figure 5: Absolute value of the GMME by model size (in millions of paramters) across four experiments: i) SIMPLE SUBJECTS (Section 4.2), ii) 2NP, iii) 3NP (Section 4.3), and iv) OBJECT CATAPHORA (Section 5.3).

6.1 Finer Comparison Between Model and Human Behavior

Following the methodology outlined in van Schijndel and Linzen (2021), we can directly compare the GMME observed in humans and in neural models. In what follows, we report on comparisons between the GMME observed for humans in Experiment 2 of Kush and Dillon (2021) and the predicted GMME in milliseconds from GPT-Neo (which was demonstrated to have non-human like behavior) and GPT-J (which did have qualitatively similar behavior to humans). To foreshadow the results, we found that both models greatly underestimate the processing cost observed in humans, even in cases of qualitative overlap.

We fit a linear-mixed effects model with reading times from the filler items in Kush and Dillon (2021) as the dependent variable, and, as fixed effects, the surprisal of the current word, the surprisal of each of the preceding three words, word length (of the current word and preceding three words), and frequency (of the current word and the preceding three words). Additional, we included fixed effects for the interaction between word length and frequency and by-participant random intercepts.[15] The predicted reading times (in milliseconds) at the subject (i.e. where we expect a GMME) were determined for GPT-Neo and GPT-J by applying the significant coefficients for the surprisal terms of their statistical model (as in van Schijndel and Linzen (2021)). For both models, the surprisal of the current word and the preceding two words were significant.[16]

Figure 4 gives the GMME for humans and the predicted GMME for the two neural models. As is visually apparent, neural models greatly underestimate the processing cost. For example, the GMME reported for humans in the condition without an interaction with Principle B was 63 milliseconds, while GPT-Neo predicted an average of around 5.7 milliseconds and GPT-J an average of around 4.7 milliseconds. 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, 2021; Wilcox et al., 2021b; Paape and Vasishth, 2022).

6.2 Model Behavior and Scale

With regards to the second remaining question, GPT-J and GPT-3 differ from the other models in one obvious way: they are the two largest models we investigated. Scaling laws suggest that larger models will outperform smaller models (e.g., Kaplan et al., 2020; Wei et al., 2022). Figure 5 plots the absolute value of the GMMEs for four of the experiments investigated in this paper, including additional results from smaller versions of GPT-2

[15] That is, we fit the model (excluding the entropy and entropy reduction terms) given in Equation 1 of van Schijndel and Linzen (2021).

[16] In particular, for GPT-Neo, the coefficients were 1.857 ms/bit for the current word, 1.802 ms/bit for the preceding word, and 1.987 ms/bit for the word two time steps in the past. Similarly, for GPT-J, the coefficients were 1.929 ms/bit, 2.037 ms/bit, and 1.980 ms/bit.

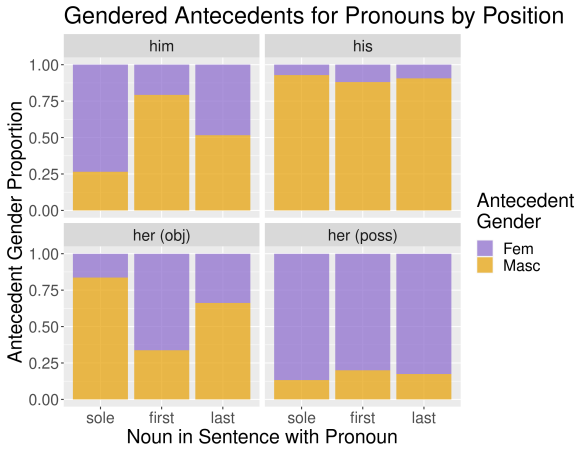

Figure 6: Proportion of each gender preceding pronouns in three positions: i) when there is exactly one antecedent, and when there is at leat two antecedents, ii) the first antecedent, and iii) the last antecedent. Data from the Pile (Gao et al., 2020) which is the training data for GPT-J and GPT-Neo.

and GPT-Neo for a larger range of model sizes. Generally, the GMME increases with model size (though GPT-3 is at times an outlier). However, for the experiment with cataphoric processing, we see that the GMME decreases with scale, suggesting that larger models learn to ignore ungrammatical positions in cataphoric pronouns, while simultaneously considering ungrammatical positions more strongly with "vanilla" pronouns.

6.3 Model Behavior and Training Data

Turning to the final remaining question (why models consider ungrammatical antecedents), the SIMPLE SUBJECT experiments are an instructive case study. Sentences like *Bill adores him* are not ungrammatical, only the interpretation that "Bill adores Bill" is blocked. Suppose the world is such the following two schema are produced at equal rates:

(12) a. Bill adores [MALE NOUN]
b. Bill adores [FEMALE NOUN]

(12-a) has two possible pronominal exponents, *Bill adores him* and *Bill adores himself*, while (12-b) has just one, *Bill adores her*. Suppose further, that the first exponent of (12-a) is twice as likely as the second. The resultant set of productions will be 50% *Bill adores her*, 33% *Bill adores him*, and 17% *Bill adores himself*.[17] Models trained on data of this sort would presumably come to favor pronouns that mismatch with the subject.

In fact, the training data for GPT-J and GPT-Neo (which is publicly available) bears resemblance to this. We took the Pile (Gao et al., 2020) and extracted all sentences with pronouns. These sentences were then parsed and chunked into noun phrases using Spacy and gender was assigned by checking for their inclusion in the male and female nouns in the BLiMP vocabulary.[18] The results are compiled in Figure 6. As is visually apparent, the data is highly indicative of a gender mismatch in the case just discussed, and skewed, to a lesser degree, towards a gender mismatch in more complex cases implicated by Principle B (e.g., 3NP stimuli).

The Binding Principles, in other words, distort the surface distribution of pronouns such that the models ultimately favor mismatches in gender in just those positions where co-indexation is impossible. Moreover, we see in the scaling figure discussed above (Figure 5), that smaller models show no, or weaker, GMMEs. Given the findings that large models have a higher capacity to memorize training data (e.g., Carlini et al., 2022; McCoy et al., 2021), we may take the GMME in the SIMPLE SUBJECT experiment to be a case of models overfiting their training data.

6.4 Conclusion

The present study argues that autoregressive models do not (uniformly) process pronouns like humans. We showed that models fail to capture the qualitative patterns of human incremental coreference processing, in addition to underestimating processing costs in constructions already noted in the literature (see van Schijndel and Linzen, 2021; Wilcox et al., 2021b). Models appear to learn only aspects of Principle B that have predictable reflexes in training data.[19] Therefore, models can mimic humans without a full human-like system. Ultimately, this work provides evidence suggesting that certain aspects of human parsing behavior do not directly follow from linguistic data. We leave bridging the gap to future work.

[17] That is, we are, for expository purposes, assuming the world consists of only structures drawn from the set {*Bill adores him*, *Bill adores her*, *Bill adores himself*} with *Bill adores herself* excluded. This is to highlight how Principle B restricts the possible strings in such a way that mismatch is more common.

[18] We used the small pretrained English model from Spacy.

[19] For a fuller discussion of mismatches between neural models and humans, as well as what these results may mean for a linguistic theory, see Davis (2022).

Acknowledgments

We would like to Dorit Abusch, Miloje Despić, Joseph Rhyne, Marten van Schijndel, Rachel Vogel, John Whitman and members of the C.Psyd lab and Cornell NLP Group, who gave feedback on earlier forms of this work. We would also like to thank the anonymous reviewers for their helpful suggestions and comments.

References

Jennifer E Arnold. 1998. *Reference Form and Discourse Patterns*. Ph.D. thesis, Stanford University.

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

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.

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.

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 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.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. arXiv.

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

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

Charles 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.

Forrest Davis. 2022. *On the Limitations of Data: Mismatches between Neural Models of Language and Humans*. Ph.D. thesis, Cornell University.

Forrest Davis and Marten van Schijndel. 2020. 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. 2021. Uncovering constraint-based behavior in neural models via targeted fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1159–1171, Online. Association for Computational Linguistics.

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.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv*.

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.

Jennifer Hu, Sherry Yong Chen, and Roger Levy. 2020. 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.

Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language Models Use Monotonicity to Assess NPI Licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969, Online. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv*.

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.

S Kennison. 2003. Comprehending the pronouns her, him, and his: Implications for theories of referential processing. *Journal of Memory and Language*, 49(3):335–352.

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 and Colin Phillips. 2014. Local anaphor licensing in an SOV language: Implications for retrieval strategies. *Frontiers in Psychology*, 5:1252.

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.

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

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.

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.

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.

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.

Jeffrey J Mitchell, Nina Kazanina, Conor J Houghton, and Jeffrey S Bowers. 2019. Do LSTMs know about Principle C? In *Proceedings of the 2019 Conference on Cognitive Computational Neuroscience*.

Janet Lee Nicol. 1988. *Coreference Processing during Sentence Comprehension*. Thesis, Massachusetts Institute of Technology.

Dario Paape and Shravan Vasishth. 2022. Estimating the true cost of garden-pathing: A computational model of latent cognitive processes. Preprint, 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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

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.

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

Hannah Rohde, Andrew Kehler, and Jeffrey L Elman. 2006. Event Structure and Discourse Coherence Biases in Pronoun Interpretation. In *28th Annual Conference of the Cognitive Science Society*, page 6.

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.

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.

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 (EMNLP)*, 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. 2021. Single-Stage Prediction Models Do Not Explain the Magnitude of Syntactic Disambiguation Difficulty . *Cognitive Science*, 45(6).

Elena Voita and Ivan Titov. 2020. Information-Theoretic Probing with Minimum Description Length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 183–196, Online. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model.

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. 2019. 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 (EMNLP-IJCNLP)*, 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. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *arXiv*.

Ethan Wilcox, Richard Futrell, and Roger Levy. 2021a. Using Computational Models to Test Syntactic Learnability. *LingBuzz*.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019. 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, 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.

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.

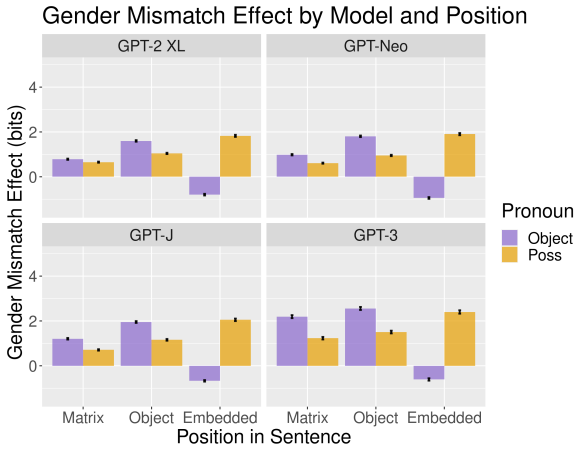

Figure 7: GMME for object pronoun (*him*) and possessive pronoun (*his*) for each neural model by whether i) the matrix subject, ii) the matrix object, or iii) the embedded subject agrees in gender (e.g., (*Bill*|*Hannah*) *told* (*Aaron*|*Amy*) that (*Mark*|*Sue*) *hates him*). Error bars are 95% confidence intervals.

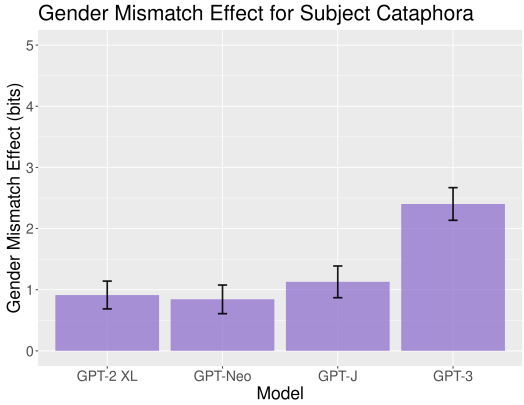

Figure 8: GMME for subject following a cataphoric subject pronoun, (e.g., *he*), for each neural model (e.g., *While he was working,* (*Bill*|*Sue*)...). Stimuli adapted from van Gompel and Liversedge (2003). Error bars are 95% confidence intervals.

A Appendix

Additional Figures

Results for the 3NP case are given in Figure 7. For the possessive pronoun *his*, we found a positive GMME for all positions, suggesting that models expected *his* to match the gender of any of the preceding antecedents. For the object pronoun *him*, a positive GMME was obtained when grammatically available antecedents (i.e. those not blocked by Principle B) mismatched in gender. A negative GMME was found for the grammatically unavailable antecedent (i.e. the embedded subject), suggesting models expected *him* to mismatch with antecedents in that structural position.

Results for subject cataphora are given in Figure 8. All models exhibited a positive GMME when the subject mismatched in gender with the cataphoric subject pronoun, suggesting that models use cataphoric subject prnouns to constrain their predictions of upcoming subjects.

Limitations

There are three main limitations: 1) whether models truly "interpret" the correct coreference relations, 2) our reliance on stereotypical gender, 3) we only investigated English.

The first was noted in Section 2. It applies to any investigation of coreference in neural models, including existing investigations of Principle A (e.g., Warstadt et al., 2020). While probing has been used to investigate model representations (e.g., Ettinger et al., 2016; Voita and Titov, 2020), which may be suggestive of something like co-indexation, we do not take models to be interpreting language, that is comprehending the meaning of sentences in a human-like fashion (see the discussion in Bender and Koller, 2020). At present, techniques are limited, and thus, we set aside the issue of whether models interpret pronouns in a human-like fashion, and instead, focus on comparing model behavior to humans, which has proved fruitful in other domains (e.g., Linzen et al., 2016). Future work might consider analyses of the attention mechanisms to dig deeper into what information models are using.

The second limitation has been noted in related literature (e.g., Warstadt et al., 2020). We rely on stereotypical associations between nouns and pronouns, which does not cleanly map on to the real world (e.g., for example, we do not consider singular *they*). In using the vocabulary items already actively manipulated in the literature, we can, nonetheless, make meaningful comparisons to existing work.

The final limitations is driven, primarily, by the existing resources in the field. There exist many pre-trained models for English, and less so for other languages (for discussion of the broader English bias in NLP, see Bender, 2009). Additional, the bulk of psycholinguistic work is focused on English, making comparisons between neural models and humans beyond English, challenging. Thus, the generalizability of the present study is limited to just those pronominal systems that are English-like.