Discourse structure interacts with reference but not syntax in neural language models

Forrest Davis and **Marten van Schijndel**
Department of Linguistics
Cornell University
{fd252|mv443}@cornell.edu

Abstract

Language models (LMs) trained on large quantities of text have been claimed to acquire abstract linguistic representations. Our work tests the robustness of these abstractions by focusing on the ability of LMs to learn interactions between different linguistic representations. In particular, we utilized stimuli from psycholinguistic studies showing that humans can condition reference (i.e. coreference resolution) and syntactic processing on the same discourse structure (implicit causality). We compared both transformer and long short-term memory LMs to find that, contrary to humans, implicit causality only influences LM behavior for reference, not syntax, despite model representations that encode the necessary discourse information. Our results further suggest that LM behavior can contradict not only learned representations of discourse but also syntactic agreement, pointing to shortcomings of standard language modeling.

1 Introduction

Neural network language models (LMs), pretrained on vast amounts of raw text, have become the dominant input to downstream tasks (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019). Commonly, these tasks involve aspects of language comprehension (or understanding). One explicit example is coreference resolution, wherein anaphora (e.g., pronouns) are linked to antecedents (e.g., nouns) requiring knowledge of syntax, semantics, and world-knowledge to match human-like comprehension.

Recent work has suggested that LMs acquire abstract, often human-like, knowledge of syntax (e.g., Gulordava et al., 2018; Futrell et al., 2018; Hu et al., 2020). Additionally, knowledge of grammatical and referential aspects linking a pronoun to its antecedent noun (reference) have been demonstrated for both transformer and long short-term memory architectures (Sorodoc et al., 2020). Humans are able to modulate both referential and syntactic comprehension given abstract linguistic knowledge (e.g., discourse structure). Contrary to humans, we find that discourse structure (at least as it pertains to implicit causality) only influences LM behavior for reference, not syntax, despite model representations that encode the necessary discourse information.

The particular discourse structure we examined is governed by implicit causality (IC) verbs (Garvey and Caramazza, 1974). Such verbs influence pronoun 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.

In addition to pronoun resolution, IC verbs also interact with relative clause (RC) attachment:

(2) a. John babysits the children of the musician who...
(i) ...lives in La Jolla.
(ii) ...are students at a private school.
b. John detests the children of the musician who...
(i) ...lives in La Jolla.
(ii) ...are arrogant and rude.
(from Rohde et al., 2011)

In (2), (2-a) and (2-b) are sentence fragments with possible continuations modifying *the musician* in (2-a-i) and (2-b-i) and continuations modifying *the children* in (2-a-ii) and (2-b-ii). We might expect human continuation preferences to be the same in (2-a) and (2-b). However, the use of an object-biased IC verb (*detests*) in (2-b) increases the proportion of continuations given by human participants that refer to the *children* (i.e. (2-b-ii) vs. (2-b-i)). Without an object-biased IC verb the majority of continuations refer to the more recent noun (i.e. *musician*).

Effects of IC have received renewed interest in the field of psycholinguistics in recent years (e.g., Kehler et al., 2008; Ferstl et al., 2011; Hartshorne and Snedeker, 2013; Hartshorne, 2014; Williams, 2020). Current accounts of IC 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 current LMs will be able to condition reference and syntactic attachment by IC verbs with just language data (i.e. without grounding).

We tested this hypothesis using unidirectional transformer and long short-term memory network (LSTM; Hochreiter and Schmidhuber, 1997) language models. We find that LSTM LMs fail to acquire a subject/object-biased IC distinction that influences reference or RC attachment. In contrast, transformers learned a representational distinction between subject-biased and object-biased IC verbs that interacts with both reference and RC attachment, but the distinction only influenced model output for reference. The apparent failure of model syntactic behavior to exhibit an IC contrast that is present in model representations raises questions about the broader capacity of LMs to display human-like linguistic knowledge.

2 Related Work

The ability of LMs to encode referential knowledge has largely been explored in the domain of coreference resolution. Prior work has suggested that LMs 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 LMs to track entities over larger spans of text (cf. Sorodoc et al., 2020). Previous work at this granularity of coreference resolution has shown LSTM LMs strongly favor reference to male entities (Jumelet et al., 2019), for which the present study finds additional support. 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 representation of referential knowledge by models trained with a general language modeling objective.

With regards to linguistic representations, a growing body of literature suggests that LSTM LMs 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., 2018), center embedding, and syntactic islands (Wilcox et al., 2018, 2019). This literature generally suggests that LMs encode some type of abstract syntactic representation (e.g., Prasad et al., 2019). Additionally, recent work has shown LMs learn linguistic representations beyond syntax, such as pragmatics and discourse structure (Jeretic et al., 2020; Schuster et al., 2020; Davis and van Schijndel, 2020a).

The robustness of these abstract linguistic representations, however, have been questioned in recent work, suggesting that learned abstractions are weaker than standardly assumed (e.g., Trask et al., 2018; van Schijndel et al., 2019; Kodner and Gupta, 2020; Davis and van Schijndel, 2020b). The present study builds on these recent developments by demonstrating the inability of LMs to utilize discourse structure in syntactic processing.

3 Language Models

We trained 25 LSTM LMs on the Wikitext-103 corpus (Merity et al., 2016) with a vocabulary constrained to the most frequent 50K words.[1] We used

[1]The models had two LSTM layers with 400 hidden units each, 400-dimensional word embeddings, a dropout rate of 0.2 and batchsize 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. The LSTMs and code for the experiments in this paper can be found at https://github.com/forrestdavis/ImplicitCausality.

two pretrained unidirectional transformer LMs: TransformerXL (Dai et al., 2019) and GPT-2 XL (Radford et al., 2019).[2]

TransformerXL was trained on Wikitext-103, like our LSTM LMs, but has 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 words compared to the 103 million in Wikitext-103). As such, we caution against extracting explicit, mechanistic claims from the present study concerning the relationship between learned linguistic knowledge and model configurations and training data. Instead, our work points to apparent differences between transformers and LSTMs with regard to use and acquisition of discourse structure, leaving explanatory principles to further work.

4 Interactions with Reference

The results of Sorodoc et al. (2020) suggested that referential contrasts based in grammatical features (e.g., gender) would be easier for models to discern then those purely focused on referential selection (e.g., antecedents with the same gender but differing in preference). To evaluate this claim, we analyzed the degree to which IC verb type (i.e. subject vs. object biased) influenced i) model pronoun preferences when the possible referents differed in gender (e.g., *Sally feared Bob because...*), and ii) similarity of model representations between the pronoun and possible referents when they share the same gender (e.g., *Fred feared Bob because he...*). Our prediction was that IC would have a weaker influence in (ii) than (i).

4.1 Referential Stimuli

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

(3) 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*).[3] 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[4] and generated stimuli as in (3) 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).[5] We created two categories of stimuli, those with differing gender[6] and those with the same gender resulting in 6888 sentences per category.

4.2 Measures

We evaluated our models independently for external behavior (e.g., predicted next-words) and internal representations (e.g., hidden states). Ideally, model behavior should condition on an abstracted representation that distinguishes subject-biased IC verbs from object-biased IC verbs. Similarly, a representational distinction between subject and object-biased IC verbs should have some influence on model behavior. We find that this is not the case for the LMs under investigation; a representational distinction between subject vs. object-biased IC verbs does not condition model behavioral differences.

To evaluate behavior, we appended a pronoun to (3) and calculated information-theoretic surprisal (Shannon, 1948; Hale, 2001; Levy, 2008). Surprisal is defined as the inverse log probability assigned to each word (w_i) in a sentence given the preceding context:

$$\text{surprisal}(w_i) = -\log p(w_i|w_1 ... w_{i-1}) \quad (1)$$

The probability of a word was calculated by applying the softmax function to a LM's output layer. Surprisal has been correlated with human processing difficulty (Smith and Levy, 2013; Frank et al., 2015) allowing us to compare model behavior to human behavior. We predicted that IC verbs would influence surprisal, with subject-biased ICs lowering the surprisal of pronouns agreeing with the subject, and object-biased ICs lowering the surprisal of the prounouns agreeing with the object. This methodology follows subject-verb agreement experiments, where verbs that agree in number with

[2] We used HuggingFace's implementation of these models (Wolf et al., 2019).

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

[4] 59 verbs were outside of our LSTM LM vocabulary, so they were excluded.

[5] See Appendix A for all the pairs.

[6] We balanced our stimuli by gender, so we had the same number of female subjects as male subjects and vice versa.

the subject are less surprising than those that do not (e.g., *Cats are* vs. **Cats is*; Linzen et al., 2016; Mueller et al., 2020)

To evaluate model representations, we followed work in representational similarity analysis and used Pearson's r[7] to measure the similarity between model representations (see Kriegeskorte et al., 2008; Chrupała and Alishahi, 2019).[8]

We build on work that has looked at the propagation of information across time within a layer (as in Giulianelli et al., 2018; Jumelet et al., 2019). In such work, model behavior for subject-verb agreement and coreference is linked to model representations, and in particular, stronger model representations of previous time steps that relate to the model's current prediction.

In the present study, we focused on the similarity between the pronoun and possible antecedents when they shared the same gender:

(4) a. The mother amused the girl because she ...
b. The mother applauded the girl because she ...

Specifically, for (4) we computed the layer-wise similarity between the hidden representation of *she* with *mother* and *girl*.[9] The bias score found in Ferstl et al. (2011) for *amused* in (4-a) was 67 (i.e. the verb is subject-biased) and for *applauded* in (4-b) it was -84 (i.e. the verb is object-biased). Thus, we predicted that a layer that encodes a human-like IC distinction should have greater similarity between *she* and *mother* in the case of *amused* than between *she* and *girl*, and vice versa for *applauded*.

[7]Specifically, corrcoef from numpy.

[8]There exist a number of other measures of representational similarity (e.g., Morcos et al., 2018). In the present study, our use of experimental materials from psycholinguistic studies resulted in far fewer data than is needed for these methods, where one wants much more data than the dimensions of the representations. This is particularly stark for the syntactic stimuli where the embedding size for GPT-2 XL is roughly 13 times larger than the number of stimuli. These techniques may ultimately provide stronger evidence for representations of implicit causality in these language models, particularly for the LSTM LMs where no representational trace of implicit causality was found. It is worth noting that the LSTM behavior does not show an influence of implicity causality, so if we were to find such a representation with a better measure of similarity it would further the disconnect between model representations and behavior we found for the transformers. We hope to explore this in future work.

[9]Given the BPE tokenizer for GPT-2 XL, if a noun was broken into components, we used the hidden representation of the final component.

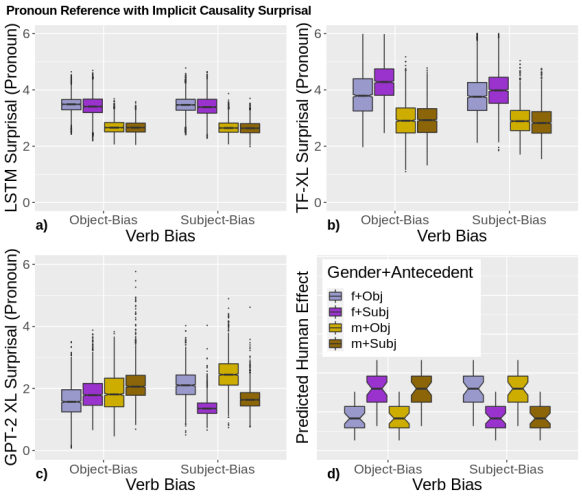

Figure 1: Model surprisal (in **a)** LSTM LMs, **b)** TransformerXL, and **c)** GPT-2 XL) at the pronoun and **d)** the predicted qualitative human-like pattern; stimuli from Ferstl et al. (2011) (e.g., *the man accused the woman because she*). Broken into antecedent (subject vs. object) and gender of pronoun (male vs. female). Lower surprisal corresponds to greater model preference.

4.3 Influence of IC on Referential Behavior

We calculated the surprisal for our LMs at the pronoun in our experimental stimuli, with the prediction that IC bias would modulate surprisal. Results for each LM type (LSTMs, TransformerXL, GPT-2 XL) are given in Figure 1. Statistical analyses[10] were conducted via linear-mixed effects models.[11] Post-hoc t-tests were conducted to assess effects.[12]

As is visually apparent in Figure 1, all three models showed some gender bias (male for TransformerXL and LSTM LMs and female for GPT-2 XL), in line with existing findings of gender preferences in LSTMs (see Jumelet et al., 2019).

The effect of IC bias was mixed across the LMs. For the LSTMs, the influence of IC was marginal ($p = 0.02$) being driven by an extremely small dif-

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

[11]We fit a model to predict surprisal at the pronoun with a three way interaction between IC bias, position of gender matching noun, and gender of pronoun and a random intercept for item. We ran a model with the continuous bias score from Ferstl et al. (2011) and another with a categorical bias effect derived from the bias score in Ferstl et al. (2011), with positive bias scores corresponding to a subject-biased verb and negative bias scores corresponding to a object-biased verb. These models had comparable results and are reported in the supplemental materials.

[12]The threshold for statistical significance was p = 0.005. Full output from the statistical models are given in the supplemental materials, and all R code to recreate the tests and figures is on Github.

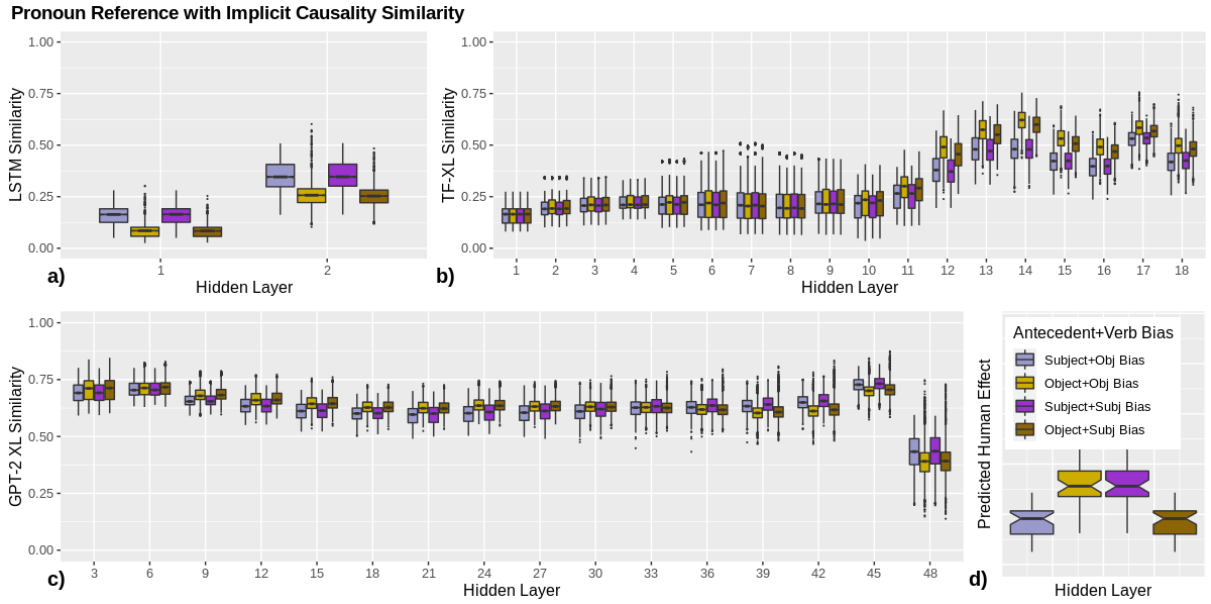

Figure 2: Layer-wise representational similarity (in **a)** LSTM LMs, **b)** TransformerXL, and **c)** GPT-2 XL) between pronoun and subject/object and **d)** the predicted qualitative human-like pattern); stimuli from Ferstl et al. (2011) (e.g., *the man accused the boy because he*). Broken into antecedent (subject vs. object) and IC bias type (subject-bias vs. object-bias). We include every third layer for GPT-2 XL (48 layers total). Greater similarity corresponds to greater relationship between pronoun and antecedent.

ference (0.02 bits) in surprisal centered on male pronouns agreeing in gender with the subject. There was neither a significant effect for object pronouns or for female pronouns referring to subjects. We concluded that the LSTM IC effect was spurious and that LSTM LMs acquired no IC-conditioned expectation about reference.

For TransformerXL, there was a slight lowering in surprisal for reference to male subjects with subject-biased verbs, and a larger lowering in surprisal for reference to female subjects after subject-biased verbs. That is to say, subject-biased IC verbs did lower the surprisal of pronouns referring to subjects, as predicted. However, there was no influence of IC when pronouns referred to the object. This suggests that preferences for local agreement in TransformerXL are much stronger than the influence of IC-bias, which only appears with subject-biased verbs.

The behavior of GPT-2 XL was in line with the human findings from Ferstl et al. (2011). Subject-biased verbs lowered the surprisal of pronouns referring to the subject, and object-biased verbs lowered the surprisal of pronouns referring to the object, regardless of gender. This suggests that GPT-2 XL has acquired a robust IC representation that influences expectations for pronominal reference.

4.4 Influence of IC on Referential Representation

We turn now to the ability of the models to distinguish the correct referent when both the subject and object have the same gender. Previous literature has suggested that this effect would be weaker than in the mismatching gender case above (see Sorodoc et al., 2020). We relied on a representational analysis (detailed in Section 4.2) to evaluate the preferences of the LMs. Results for each LM type are given in Figure 2.

Statistical significance was determined via linear-mixed effects models with post-hoc t-tests assessing the effects.[13] As predicted, IC bias had a weaker effect when choosing between competing nouns with the same gender for reference (e.g., *the woman admires the queen because she*).

For LSTM LMs, IC bias did not influence model representations, at least as measured in the present study. For TransformerXL there was a small difference in degree of similarity from layers 12 to

[13]We fit models predicting similarity from a four way interaction of IC bias, noun comparison (subject or object), model layer, and gender and a random intercept for item. Two IC bias effects were considered: the gradient value given in Ferstl et al. (2011) and a categorical value where positive bias corresponds to a subject-biased verb and negative bias corresponds to a object-biased verb. Similar results were found with both effects with both models given in the supplemental materials.

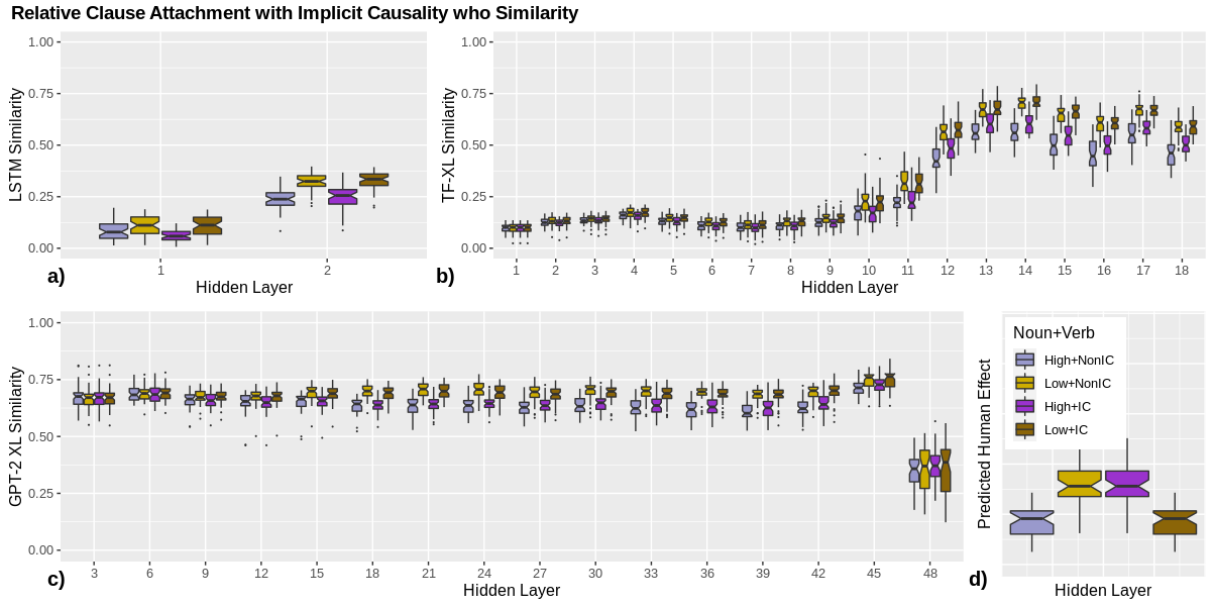

Figure 3: Layer-wise representational similarity (in **a)** LSTM LMs, **b)** TransformerXL, and **c)** GPT-2 XL) between *who* and the higher/lower noun, and **d)** the qualitative predicted human-like pattern); stimuli from Rohde et al. (2011) (e.g., *the man admired the agent of the rockers who*). Broken into attachment location (higher noun vs. lower noun) and verb type (object-biased IC verb vs. non-IC verb). We include every third layer for GPT-2 XL (48 layers total). Greater similarity corresponds to greater relationship between attachment location and *who*.

18. The pronoun was more similar to the object when the verb was object-biased. In contrast, there was no significant effect for subject-biased verbs, despite the reverse effect in behavior when antecedents had mismatched gender (i.e. subject-bias, not object-bias, influenced pronoun surprisal in our behavioral analysis).

For GPT-2 XL we found a small, yet significant, difference in degree of similarity with the subject antecedent starting in layer 15 and continuing through layer 47. That is, there was greater similarity between the pronoun and the subject when the verb was subject-biased. There was no effect for pronouns referring to the object. These results suggest that the influence of IC is only weakly present when both the subject and object are possible antecedents (i.e. they are the same gender). It therefore seems that models were only able to fully leverage an IC contrast to resolve reference when gender differences unambiguously distinguished between subject and object.

5 Interactions with Syntactic Attachment

We turn now to the relationship between IC verbs and syntax. Recall the prediction that object-biased IC verbs should interact with RC attachment to license more cases of syntactic attachment to the higher noun compared to lower noun (i.e. *chef* in *Anna scolded the chef of the aristocrats who was/were...*).

5.1 Syntactic Stimuli

We used stimuli from Rohde et al. (2011), which consisted of two experiments: a sentence completion task and a self-paced reading task. The sentence completion task consisted of 21 prompts like:

(5) a. Carl admires the agent of the rockstars who...
b. Carl works with the agent of the rockstars who...

The key manipulation lies with the main verb. In (5-a), *admires* is an object-biased IC verb, and in (5-b), *works with* is a non-IC verb. In the present study, we took 14 of these prompts[14] and generated stimuli balanced for number (i.e. we added *Carl admires the agents of the rockstar who...*), for a total of 112 sentences.

The self-paced reading time study in Rohde et al. (2011) consisted of 20 pairs of sentences, as in:

(6) a. Anna scolded the chef of the aristocrats who was/were routinely letting

[14] 7 prompts were excluded because either the non-IC or the IC verb was not in the vocabulary of our LSTM LMs. For the remaining prompts, we replaced ceo(s) with boss(es), supermodel(s) with superstar(s), and rockstar(s) with rocker(s).

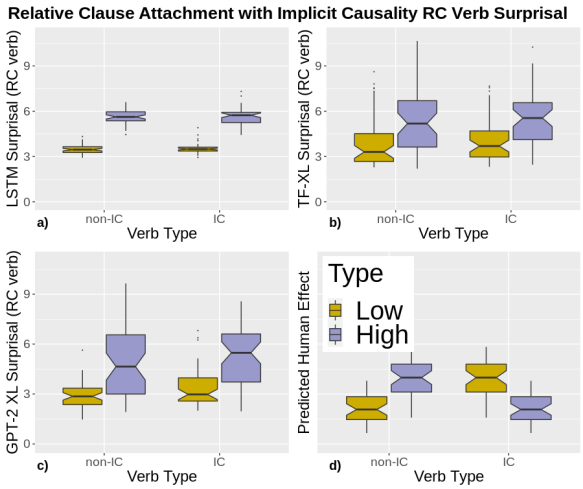

Figure 4: Model surprisal (in **a)** LSTM LMs, **b)** TransformerXL, **c)** GPT-2 XL, and **d)** qualitative predicted human-like pattern) at the RC verb (*was/were*); stimuli from Rohde et al. (2011) (e.g., *the man admired the agent of the rockers who was/were*). Broken into location of agreement (High vs. Low). Lower surprisal corresponds to greater model preference.

food go to waste.

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

As with the completion study, the central manipulation in the self-paced reading study lies with whether the verb is an object-biased IC verb (*scolded*) or not (*studied with*). Rather than give completions, though, 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 again 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.[15]

[15]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). Given that all our LMs were unidirectional, we ignored the material after the RC verb. Additionally, for both the completion and self-paced reading stimuli, we substituted male names with *the man* and female names with *the woman*.

5.2 Measures

For the sentence completion stimuli, we conducted a cloze task. Specifically, given the sentence fragment *the man admires the agent of the rockstars who*, we calculated the top 100 most likely next words for the LMs. These were then tagged for part-of-speech using spaCy and a score was assigned based on the weighted probability of a continuation using a singular verb (i.e. probability mass assigned to singular verbs divided by probability mass assigned to all verbs).[16] Our prediction is that object-biased IC main verbs will lead to more continuations agreeing with the higher noun (e.g., *agent*).

As detailed in Section 4.2, we calculated information-theoretic surprisal and layer-wise similarity. With the self-paced reading time stimuli (e.g., 6), we calculated surprisal at the RC verb and calculated similarity between *who* (and *was/were*) and the higher and lower nouns (e.g., *chef* and *aristocrats* for (6)). We predicted that with object-biased IC verbs like *scolded* in (6) there would be greater similarity between *who* and *chef* than for *who* and *aristocrats* in layers that have an IC distinction (vice versa for non-IC verbs like *studied*).

5.3 Influence of IC on Syntactic Behavior

To test the influence of IC verbs on model behavior for RC attachment, we followed the experiments in Rohde et al. (2011). The results are given in Figure 4. We evaluated statistical significance with linear-mixed effects models.[17] Post-hoc t-tests were conducted to assess any effects. None of the LM architectures showed any influence of IC on model behavior, either for the cloze task or the self-paced reading stimuli. Rather, they all had a strong preference for agreeing with the lower noun.[18]

[16]We excluded verbs that were ambiguous (e.g., *ate*).

[17]We fit models predicting surprisal at the RC verb from a three way interaction of agreement location, main verb type (object-biased IC or not), and number and a random intercept for item. For the cloze task, we fit models predicting percent singular continuation from an interaction between location of singular agreement (higher or lower noun) and main verb type and a random intercept for item.

[18]With regard to categorical preferences (i.e. numerically lower surprisal for one attachment location over another for a given stimulus), all the LMs have overwhelming preferences for attachment to the lower noun. The LSTM LMs favored attachment to the higher noun in 0% of stimuli (across both IC and non-IC stimuli). For TransformerXL, attachment to the higher noun was preferred in 25% of the stimuli with an object-biased IC verb (i.e. where we expect a preference for attachment to the higher noun) and attachment to the higher noun in 27% of the stimuli without an object-biased IC verb

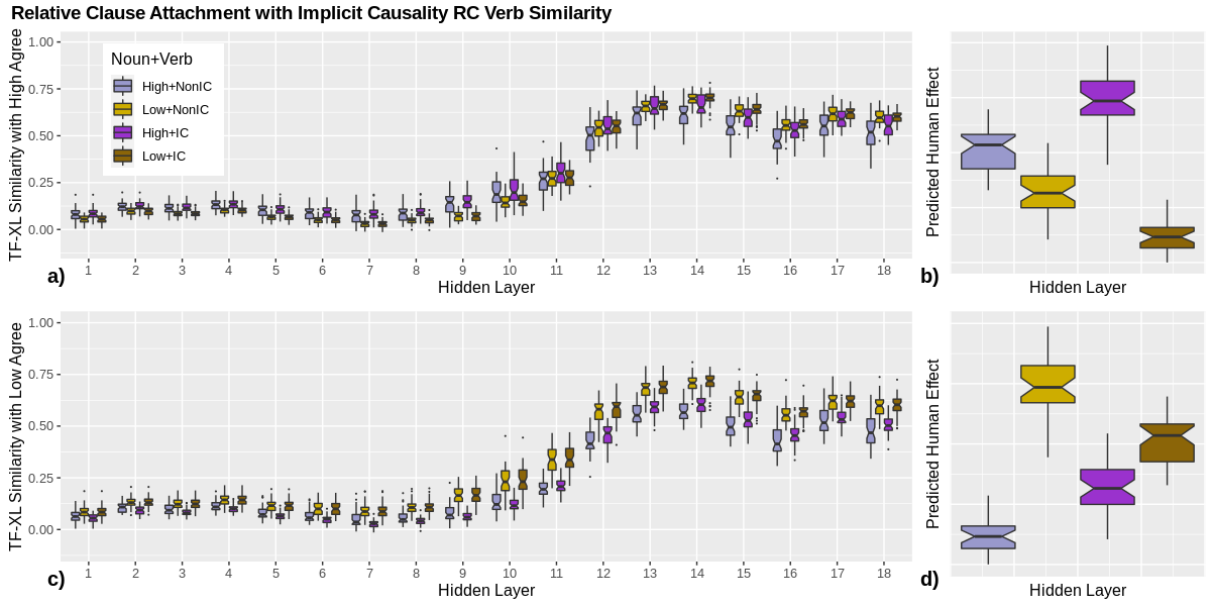

Figure 5: Layer-wise representational similarity between the RC verb (*was/were*) and the higher/lower noun; stimuli from Rohde et al. (2011) (e.g., *the man admired the agent of the rockers who was/were*). Results broken into attachment location (higher noun vs. lower noun) and verb type (object-biased IC verb vs. non-IC verb) are given in **a)**, for stimuli where the RC verb agrees with the higher noun (e.g., *agent of the rockers who was*), and in **c)**, for stimuli where the RC verb agrees with the lower noun (e.g., *rockers who were*). The explicit agreement should force a particular attachment location to be preferred, with verb IC bias dampening this effect (the predicted qualitative human-like pattern is depicted in **b)** and **d)**). Greater similarity corresponds to greater relationship between attachment location and *was/were*.

5.4 Influence of IC on Syntactic Representations

We examined the representational similarity between *who* and the possible attachment points (i.e. higher or lower noun) and the RC verb (*was/were*) and the possible attachment points. Results for all three LM architectures for *who* are given in Figure 3, and results for the RC verb are given for TransformerXL in Figure 5. Statistical significance was determined via linear-mixed effects models.[19] Post-hoc t-tests were conducted to assess effects.

The LSTM LMs had no representational effect of IC on either *who* or the RC verb, similar to the lack of an effect in pronouns. Instead, the LSTM LMs representations always had greater similarity to the lower noun, in line with the robust preference for attaching to the lower noun in behavior (i.e. model surprisal).

For TransformerXL, object-biased IC verbs increased the similarity between the higher noun and both *who* and the RC verb (regardless of agreement). That is, the presence of an object-biased IC verb increased the similarity of the RC verb and the higher noun both when the RC verb agreed in number with the higher noun (e.g., *chef...was*) and when the RC verb did not agree in number (e.g., *chef...were*). There was no effect of IC on the similarity between the lower noun and *who* or the RC verb.

For GPT-2, object-biased IC verbs increased the similarity between the higher noun and *who*, but only increased the similarity between the RC verb and the higher noun when they agreed in number (i.e. increased similarity between *chef* and *was* not *chef* and *were*). As with TransformerXL, there was no analogous effect on the similarity between the lower noun and *who* or the RC verb (i.e. no change in similarity based on the main verb).

We found TransformerXL had greater similar-

(i.e. where we do not expect attachment to the higher noun). Finally, for GPT-2 XL, for the stimuli with an object-biased IC verb attachment to the higher noun was preferred in 23% of the stimuli, and for stimuli without an object-biased IC verbs 9% of the time.

[19] Specifically, we fit a model predicting the similarity between *who* and the possible nouns with a three way interaction of main verb type (object-biased IC verb or not), noun (higher or lower), and layer with a random intercept for item. Additionally, we fit a model predicting the similarity between *was/were* and the possible nouns with a four way interaction of main verb type (object-biased IC verb or not), noun (higher or lower), agreement location, and layer with a random intercept for item.

ity between the RC verb and the lower noun in the final layer, regardless of verbal agreement (i.e. even in cases of ungrammatical attachment, TransformerXL preferred local attachment). Similarly, GPT-2 XL showed no preference for attachment location in the final layers despite unambiguous agreement with only one of the nouns. Strikingly, both transformer LMs showed greater similarity with the agreeing noun (i.e. similarity conditioned on syntax) in their earlier layers, with the final layers obscuring this distinction.

These results suggest that a preference for local agreement is robust in both LSTMs and transformer LMs. The transformers showed representations that encoded the IC contrast, as with the referential experiments. However, this knowledge did not propagate to the final layers, in line with the absent behavioral effects detailed above. Moreover, unambiguous syntactic knowledge about RC attachment was discarded in the final layers of TransformerXL and GPT-2. These results suggest that non-linguistic locality preferences dominate model representations and behavior.

6 Discussion

The present study examined the extent to which discourse structure, determined by implicit causality verbs, could be acquired by transformer and LSTM language models (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 reference and syntactic attachment in LMs. Analyses were conducted at two levels of granularity: model behavior (e.g., probability assigned to possible next words) and model representation (e.g., similarity between hidden representations). Given the claims in recent literature that implicit causality arises without extra pragmatic inference on the part of human comprehenders, we hypothesized that LMs would be able to acquire such contrasts (analogous to their ability to acquire syntactic agreement).

We found that LSTM LMs were unable to demonstrate knowledge of IC either in influencing reference or syntax. However, a transformer (TransformerXL) trained on the exact same data as the LSTM LMs was able to partially represent an IC distinction, but model output was only influenced by IC bias when resolving reference, not syntactic attachment. In evaluating a transformer model trained on vastly more data (GPT-2 XL), we found a more robust, human-like sensitivity to IC bias when resolving reference: subject-biased IC verbs increased model preference for subject pronouns and object-biased IC verbs increased model preferences for object pronouns. However, the same mismatch as TransformerXL between model representation and model behavior arose in processing syntactic attachment.

In contrast to our results, Davis and van Schijndel (2020a) 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. The erasure of number agreement in the final layers of the transformer LMs (see Section 5.4) provides compelling evidence towards this conclusion.[20]

From a theoretical perspective, the present study provides 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 1 (see also Hartshorne, 2014). The mismatches in syntactic representations and behavior suggest, however, that models ignore the abstract categories that are learned, contrary to human findings (cf. Rohde et al., 2011).

We believe a solution may lie in changing model training objectives (i.e. what linguistic unit should be predicted). Psycholinguistic studies focusing on the interaction of discourse and syntax have suggested that coherence relations may be the unit of linguistic prediction, in contrast to the next-word prediction used in most language modeling work (see Rohde et al., 2011). We leave to future work an investigation of this suggestion as well as teasing apart the exact role that training data and model architecture play in the interaction between types of linguistic representation.

Acknowledgments

Thank you to members of the C.Psyd lab at Cornell, who gave feedback on an earlier form of this work. We would also like to thank the three anonymous reviewers for their comments and suggestions.

[20] Further cross-linguistic evidence bearing on the inability of LSTM LMs, specifically, to learn relative clause attachment is given in Davis and van Schijndel (2020b).

References

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.

Jean-Philippe Bernardy and Shalom Lappin. 2017. Using deep neural networks to learn syntactic agreement. *Linguistic Issues in Language Technology*, 15.

Pengxiang Cheng and Katrin Erk. 2020. Attending to Entities for Better Text Understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34.

Grzegorz Chrupała and Afra Alishahi. 2019. Correlating neural and symbolic representations of language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2952–2962, Florence, Italy. 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. Interaction with Context During Recurrent Neural Network Sentence Processing. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*.

Forrest Davis and Marten van Schijndel. 2020b. 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.

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 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

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 (CoNLL 2017)*, pages 3–14. Association for Computational Linguistics.

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.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain & Language*, 140:1–11.

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

Catherine Garvey and Alfonso Caramazza. 1974. Implicit causality in verbs. *Linguistic inquiry*, 5(3):459–464.

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.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. 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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. 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.

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, 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. 2008. Coherence and coreference revisited. *Journal of semantics*, 25(1):1–44.

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.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

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*, 4:521–535.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Wikitext-103. Technical report, Salesforce.

Ari Morcos, Maithra Raghu, and Samy Bengio. 2018. Insights on representational similarity in neural networks with canonical correlation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5727–5736. Curran Associates, Inc.

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.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 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*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI.

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

Hannah Rohde, Roger Levy, and Andrew Kehler. 2011. Anticipating explanations in relative clause processing. *Cognition*, 118(3):339–358.

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.

Claude Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

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.

Andrew Trask, Felix Hill, Scott E Reed, Jack Rae, Chris Dyer, and Phil Blunsom. 2018. Neural arithmetic logic units. In *Advances in Neural Information Processing Systems*, pages 8035–8044.

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*. Association for Computational Linguistics.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2018. What Syntactic Structures block Dependencies in RNN Language Models? In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*.

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*.

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, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

A Stereotypically gendered nouns used in referential experiments

male	**female**
man	woman
boy	girl
father	mother
uncle	aunt
husband	wife
actor	actress
prince	princess
waiter	waitress
lord	lady
king	queen
son	daughter
nephew	niece
brother	sister
grandfather	grandmother