The Role of Physical Inference in Pronoun Resolution

Cameron R. Jones (c8jones@ucsd.edu)
Department of Cognitive Science, UC San Diego,
9500 Gilman Dr., La Jolla, CA 92093, USA

Benjamin K. Bergen (bkbergen@ucsd.edu)
Department of Cognitive Science, UC San Diego,
9500 Gilman Dr., La Jolla, CA 92093, USA

Abstract

When do people use knowledge about the world in order to comprehend language? We asked whether pronoun resolution decisions are influenced by knowledge about physical plausibility. Results showed that referents which are more physically plausible in described events were more likely to be selected as antecedents of ambiguous pronouns, implying that resolution decisions were driven by physical inference. An alternative explanation is that these decisions were driven instead by distributional word knowledge. We tested this by including predictions of a statistical language model (BERT) and found that physical plausibility explained variance on top of the statistical language model predictions. This indicates that at least part of people's pronoun resolution judgments comes from knowledge about the world and not the word. This result constrains psycholinguistic models of comprehension—world knowledge must influence propositional interpretation—and raises the broader question of how non-linguistic physical inference processes are incorporated during comprehension.

Keywords: pronoun interpretation; language comprehension; world knowledge; situation models; language models

Introduction

Theories of language comprehension vary in the role they assign to world knowledge in determining interpretation. In particular, theories differ in whether they allow world knowledge to influence the propositional interpretation of an utterance. Some theories assign a **minimal** or **elaborative** role, where world knowledge is not integrated before a propositional interpretation has been generated. Other theories provide a mechanism for world knowledge to influence propositional interpretation, either by **validating** a given interpretation against world knowledge, or using world knowledge to generate **expectations** before interpretation begins.

Imagine a reader encounters the following sentence:

(1) The swallow carried the coconut.

Traditional situation model theories of language argue that a comprehender transforms this input into propositions, which form the basis of their representation of the message (Kintsch & Van Dijk, 1978). For instance (1) might be transformed into the proposition *carry(swallow, coconut)*. According to McKoon & Ratcliff's *Minimalist Hypothesis* (1995, 2015), world knowledge will play a **minimal** role in interpreting such input. They argue that inferences are only made when supporting information is "highly available" or there is a break in local coherence which triggers strategic analysis. As neither condition is met, there ought to be no further inferences made.

Kintsch's (1998) *Construction-Integration* model supports a more pervasive (though still limited) influence of world knowledge. Input is transformed into propositions which activate related propositions of world knowledge in a pre-existing semantic network. Activation then spreads throughout the network until an equilibrium state is reached. The model allows related information to be integrated into the comprehender's mental model, **elaborating** on the propositional interpretation of the input. For instance, a comprehender of (1) might activate information about the shapes of swallows and coconuts and infer that the swallow grips the coconut's husk with its claws. Although the Construction-Integration model offers more potential for elaborative inference than the Minimalist Hypothesis, it is crucially similar in that a propositional interpretation of the input is selected before world knowledge can be activated.[1] Thus world knowledge can play only an elaborative role in supporting the core propositional information of the linguistic signal.

Other theories of language comprehension provide mechanisms for world knowledge to influence the core propositional interpretation of linguistic input: either by rejecting an implausible interpretation (validation) or by influencing the comprehension process before a propositional interpretation has been chosen (expectation). O'Brien & Cook's 3-stage *RI-Val* model (2016) is similar to the Construction-Integration model in its first two stages (*Resonance* and *Integration*). The third stage, *Validation*, checks the output against the existing situation model and general world knowledge. If an inconsistency is discovered, the initial propositional interpretation may be rejected or revised. For instance, a reader of (1) might initially interpret *swallow* as referring to the small European Swallow. During **validation** they might conclude that the described situation is implausible, and reinterpret *swallow* as referring to the larger African variant.

Sanford and Garrod's (1998) *Scenario-Mapping* model is a more radical departure from traditional situation model theories. They propose that linguistic input is integrated with background knowledge immediately, before selecting a propositional interpretation for the input. This allows world knowledge to influence interpretation at the earliest possible opportunity by generating **expectations** about the proposi-

[1] Although the C-I model provides a mechanism for world knowledge to winnow potential interpretations during integration, the process is heavily influenced by which interpretations are initially selected and world knowledge plays no role in this initial selection.

tional form which input must take.

In the former two models, world knowledge cannot typically affect the propositional interpretation of input, but on the latter two it can. If world knowledge does indeed play a role in determining propositional interpretation, this would have implications for a variety of areas of cognitive science. It would place constraints on psycholinguistic models of comprehension, which would need to account for the rapid access and influence of such information. It would suggest that natural language understanding solutions will need to incorporate world knowledge in order to achieve human performance on a variety of tasks. Finally, it would provide evidence for the continuity of language processing and other cognitive tasks, and raise questions about how cognitive resources are shared between linguistic and non-linguistic activities.

A variety of research provides supportive but inconclusive evidence on the specific question of whether world knowledge influences the propositional interpretation of linguistic input. Online measures such as Event-Related Potentials and reading times provide invaluable insight into the time-course of world knowledge activation. World knowledge violations cause early processing difficulty, indicating that world knowledge must be activated and integrated rapidly during comprehension (Hagoort, Hald, Bastiaansen, & Petersson, 2004; Garrod, Freudenthal, & Boyle, 1994; Milburn, Warren, & Dickey, 2016). However, online measures alone cannot expose the result of such integration: the products of language comprehension (Ferreira & Yang, 2019). Additional evidence is needed using methods which probe the contents of mental representations during and after comprehension.

Visual world studies provide evidence that world knowledge can influence anticipation of upcoming linguistic input, which in turn drives eye-movement behavior (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Altmann & Kamide, 1999). However, the presence of visual stimuli before and during processing may facilitate exceptionally strong world knowledge influence that might not generalise to ordinary reading behavior or comprehending spoken language in the absence of visual referents.

Ambiguous language provides an ideal testbed for the influence of world knowledge on comprehension. If a linguistic signal affords multiple propositional interpretations, and world knowledge influences the interpretation which a comprehender selects, we should expect that comprehenders disproportionately select world knowledge-consistent interpretations, controlling for other factors. Related research uses ambiguous pronouns to investigate the implicit causality of verbs: the tendency of comprehenders to attribute causal responsibility for the event described by a verb to a particular entity. Comprehenders resolve *she* to *Sally* in *Jane criticized Sally because she...*, but to *Jane* if the verb is changed to *amazed* (Garvey & Caramazza, 1974). Although some researchers have interpreted implicit causality effects as the influence of world knowledge about the typical causes of events (Van den Hoven & Ferstl, 2018; Pickering & Majid, 2007), others have argued that they result from purely linguistic knowledge about verbs (Hartshorne, 2014).

Two studies have explicitly examined the influence of world knowledge on pronoun interpretation. As part of a pilot study for a self-paced reading experiment, Gordon & Scearce (1995) found that pronoun interpretation is influenced by modulating the verb in sentences like *Bill wanted John to look over some important papers... Unfortunately he never [sent/received] them*. More recently, Bender (2015) found that human subjects perform well on the Winograd Schema Challenge, an artificial intelligence benchmark that involves solving pronoun resolution problems designed to require world knowledge. Both studies face two methodological challenges, which we aim to address in the present study.

First, both studies use the experimenter's intuition as their metric of world knowledge plausibility. Experimenters' judgements might not align with participants', or may have been guided by pragmatic, lexical, or other non-world knowledge information that is known to influence pronoun resolution judgements. We therefore conducted two separate norming studies to measure i) the world knowledge bias of each of our stimuli, and ii) the bias exerted by the structural linguistic information in the stimulus, in the absence of world knowledge. These measures not only provide an experimenter-independent metric of world knowledge bias, but also allow us to control for the effects of structural information. Moreover, by operationalizing the strength of each bias as the proportion of participants who chose a particular response, we can treat both biases as continuous variables. This allows us to test the stronger claim that the degree of bias should predict the degree of effect on responses.

Second, many previous experiments have not controlled for the possibility that participants use their distributional knowledge of language to provide responses that are consistent with world knowledge (Willits, Amato, & MacDonald, 2015). We test this account by measuring the variance explained by world knowledge biases when controlling for the predictions of a language model. If participants are using distributional information, rather than experiential world knowledge, to make pronoun resolution judgements, we expect a language model (which makes use of this same distributional information) to explain any variance in their responses which might be explained by world knowledge bias.

The Present Study

In order to investigate the role of world knowledge in language comprehension, we test whether knowledge about the physical world exerts an influence on pronoun interpretation. When a comprehender encounters the pronoun *it* in (2), they must decide whether it refers to *the vase* or *the rock*.

(2) When the vase fell on the rock, it broke.

Linguists have found evidence for a variety of structural factors which influence pronoun interpretation. Comprehenders are more likely to resolve a pronoun to the subject of the previous clause (Crawley, Stevenson, & Kleinman, 1990), or to a

Table 1: Item versions and responses

Study	Order	Stimulus	NP2 Responses
Structural Norming	A	When the purple vase fell on **the green vase**, it broke.	10%
Structural Norming	B	When the green vase fell on **the purple vase**, it broke.	0%
Physics Norming	A	If a vase fell on **a rock**, which would be more likely to break?	0%
Physics Norming	B	If a rock fell on **a vase**, which would be more likely to break?	100%
Main Experiment	A	When the vase fell on **the rock**, it broke.	12.5%
Main Experiment	B	When the rock fell on **the vase**, it broke.	95%

noun phrase (NP) which occupies the same grammatical role as the pronoun (Chambers & Smyth, 1998). In (2), these factors both bias interpretation toward *the vase*. Orthogonally, if comprehenders are using world knowledge to assess the plausibility of candidates, then all things being equal, they should select antecedents that they think are more plausible in the described events. Consider:

(3) When the rock fell on the vase, it broke.

Here, structural factors continue to bias interpretation toward the first noun phrase (NP1—now, *the rock*). However, world knowledge about the physical properties of rocks and vases may lead an interpreter to believe this interpretation is implausible and select an alternative antecedent: *the vase* (Hobbs, 1979; Garnham, 2001). Thus if world knowledge does influence interpretation, more participants should choose the second noun phrase (NP2) in (3) than in (2).

A difference in NP2 responses between (2) and (3) would allow us to reject the null hypothesis that *only* structural factors influence pronoun assignment. However, it does not entail that world knowledge is responsible for this shift. Therefore, we conducted separate norming studies measuring the strength of structural and physics biases. We elicited structural norms by replacing the NPs in each experimental item with two NPs deemed equally likely to participate in the critical event (see Table 1, rows 1-2). We elicited physical norms by asking participants explicit hypothetical reasoning questions about the described situation (Table 1, rows 3-4). We operationalize structural and physics biases as the proportion of participants who selected NP2 for each item in the respective norming studies.

In addition to structural biases, participant decisions may be influenced by statistical associations between words. A participant might select *the vase* in (3) purely because *broke* is more likely to follow *the vase* than *the rock*. Therefore, a comprehender could in principle use statistical knowledge about the distribution of language to produce responses which are consistent with their physical knowledge. This could result from the physical world influencing language production: language is often used to describe the real world, and so the relevant world knowledge may be encoded in the distribution of language itself. Alternatively, distributional information may not fully capture the ground truth of the physical world, but linguistic knowledge might influence comprehenders' interpretation even when it conflicts with knowledge gleaned from their grounded experience with the physical world. In order to control for the influence of distributional linguistic knowledge, we also obtained probabilities for each antecedent from a language model (LM). LMs learn to predict sequences of words based on the statistical distribution of words in language. If LM predictions account for the variance in participant responses that is explained by world knowledge bias, this supports an alternative explanation that does not require the influence of world knowledge.

Different theories of world knowledge make distinct predictions about the effect of these measures on pronoun resolution judgements. Models which assign only a **minimal** or **elaborative** role predict no marginal effect of physics bias. World knowledge, on these accounts, exerts no influence before a propositional interpretation has been extracted. It can therefore play no role in determining the propositional interpretation. In contrast, accounts which assign a role to world knowledge in **validation** or generating **expectations** do predict a marginal effect of physics bias. Validation accounts predict that physically implausible referents, such as *the rock* in (3) will be ultimately rejected, even if they are favoured by structural factors. Similarly, if world knowledge influences **expectations** about how the text will develop, then *the vase* might be preferred initially, without the need for alternate candidates to be validated.

Method

Norming Studies

Participants All research was approved by the University of California San Diego Institutional Review Board. We recruited 35 native English speaking undergraduate students from the Psychology Deparment Subject pool, who provided informed consent using a button press and received course credit as compensation for their time. All participants successfully answered $\geq 2/3$ catch trials. We excluded 1 participant who indicated they were not a native English speaker; 1 participant who took over 1 hour to complete the experiment; and 8 participants who had $> 20\%$ of their trials excluded. We excluded 43 trials where the response time was $< 500ms$ (indicating guessing), and 67 trials where the response time (offset by 191ms per syllable of question length in the syntax condition) was $> 10s$ (indicating inattention or deliberation). We also excluded 5 trials where the response time

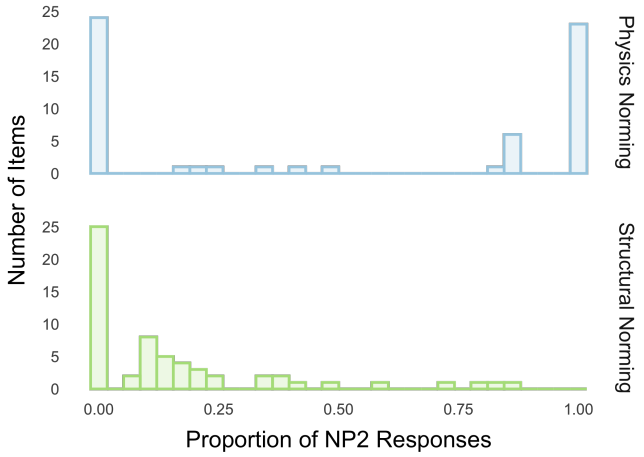

Figure 1: Distribution of biases elicited via norming. Physics bias is bimodally and symmetrically distributed while the structural biases are unimodally skewed toward NP1.

was $\pm 2.5SD$ from the participant mean. 1 participant used a touchscreen and none of their trials met exclusion criteria. We retained 820 trials (375 physics, 445 structural) from 27 participants (13 physics, 14 structural; 20 female, 6 male, 1 non-binary; mean age = 20.3, $\sigma = 1.8$). The physics norming study lasted 6.9 mins on average ($\sigma = 1.3$), while the structural norming lasted 19.1 mins on average ($\sigma = 4.2$). The difference in duration was largely due to the inclusion of filler items in the structural norming study.

Materials We created two alternate versions of each of the critical items from the main experiment (see § Main Experiment). To elicit structural norms, we replaced the candidate antecedents with two objects which were deemed to be equally physically plausible. We used either modifiers that did not alter the physical properties relevant to the plausibility of the candidate, or different objects that were similar in relevant properties. We confirmed that differences between the objects were not influencing decisions by calculating the difference in proportion of NP2 responses when the order of the NPs was reversed ($\mu = 0.1$). To elicit physics norms we reframed the pronoun resolution problem as an explicit reasoning task (see Table 1). [2]

Procedure The experiment was designed using jsPsych (De Leeuw, 2015) and hosted online. Passages were presented for $250ms + 191ms/syllable$. A question would then appear below the sentence along with two response options. In the physics norming study, the question was presented immediately and the response options were revealed after a delay. Participants were instructed to use the keyboard (or touchscreen) to indicate their response. Two examples were then presented, along with instructions on how to respond in each case. The examples were counterbalanced with respect to presentation order, and (in the structural norming) did not require the use of physical inference to resolve.

Participants in both norming tasks were presented with 30 critical items and 3 catch trials. In the structural norming study, 45 filler items were included in order to mask the purpose of the study from participants. Filler items were taken from other studies about pronoun resolution (Bender, 2015; Crawley et al., 1990; Smyth, 1994). Fillers were vetted to ensure they did not encourage physical inference and balanced with respect to NP1/NP2 bias. Presentation order of items was randomized. The position of response options was also randomized, so that the NP1 response appears on the right in half of trials.

Results Responses were aggregated by item to find the proportion of NP2 responses in each norming study. Results for a single item are shown in Table 1. Items in the structural norming study elicited responses which were heavily skewed toward NP1 (see Figure 1). This is likely due to subjecthood biases (as NP1 was often the subject) and grammatical parallelism (as ambiguous pronouns were often grammatical subjects). Most responses in the physics norming study elicited 0% or 100% NP2 responses, indicating high agreement and reflecting the fact that reversing the order of each item effectively reverses its bias with respect to NP1/NP2-coding.

Main Experiment

Participants Participants were recruited, excluded, and compensated in the same manner as described for the norming studies. 48 participants were recruited, and 15 were excluded (5 non-native English; 1 failed $\geq 2/3$ catch trials; 1 with completion time $>$ 1hr; 8 with $>$ 20% trials excluded) leaving 33 (20 female, 10 male, 1 non-binary, 2 prefer not to say; mean age = 20.3, $\sigma = 2.6$). Mean completion time was 18.8 minutes ($\sigma = 5.0$). We excluded trials where response time was $< 500ms$ (46), $> 10s$ ($+191ms/syllable$, 105), $\pm 2.5SD$ from participant mean (5), leaving 1105 trials.

Materials 30 critical items were designed so that each featured an introductory clause that referred to two objects (the candidates), and an ambiguous pronoun that referred back to one of the candidates in a later clause. The later clause described a physical event in which one of the candidates was a more plausible participant than the other, such as in (2). We used a variety of situations, which would require invoking different physical properties to infer the most plausible candidate, including mass, velocity, momentum, brittleness, mass distribution, surface area, scratch hardness, indentation hardness, melting point, and flammability. All items were designed so that the candidates could be switched and the order of the candidates was randomized across participants, forming pairs (see Table 1, rows 5-6).

Procedure The main experiment proceeded exactly as the structural norming study, described above (including the same instructions and filler items).

[2] All items, as well as the source code for the experiment, are available on github: https://github.com/camrobjones/pipr.

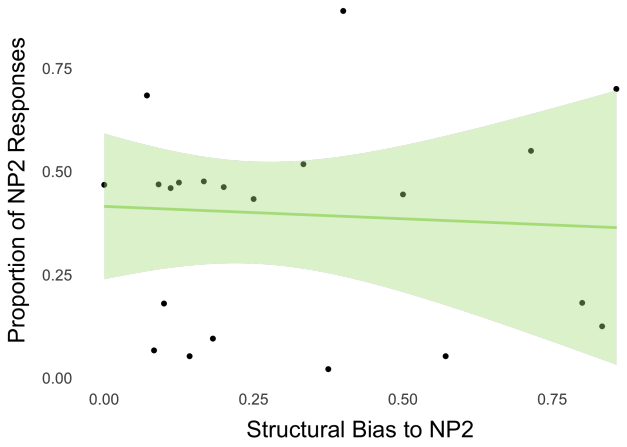

Figure 2: Structural factors, such as grammatical role, had little influence on whether an NP was selected as an antecedent, $r = -0.09$, $\chi^2(1) = 0.407$, $p = 0.52$.

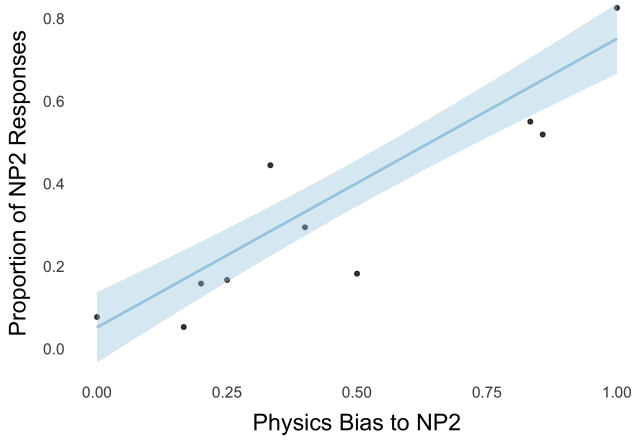

Figure 3: The physical plausibility of an NP had a strong effect on whether it is selected as an antecedent, $r = 0.72$, $\chi^2(1) = 65.667$, $p < 0.001$.

Results

Main Experiment

We constructed linear mixed effects models using the *lme4* package in R (Bates, Sarkar, Bates, & Matrix, 2007). Models predicted the responses in the main experiment, with random effects of physics and structural bias by participant. We used Likelihood Ratio Tests to compare models. No significant effect of structural bias was detected compared to a null model ($\chi^2(1) = 0.407$, $p = 0.52$; marginal $R^2 < 0.001$; see Figure 2). However, a model which included physics bias performed significantly better than a model with only structural bias as a fixed effect ($\chi^2(1) = 65.667$, $p < 0.001$; marginal $R^2 = 0.55$; see Figure 3). Thus, physics bias appears to have a strong positive effect on pronoun resolution in the main experiment, while there is no clear effect of structural bias.

Language Model Analysis

The significant effect of physics bias observed above could represent the causal influence of world knowledge in pronoun resolution. However, this result is also consistent with an alternative account: participants may be using statistical knowledge about the distribution of language to select an antecedent. Language is often generated to describe the real world and so we would expect judgements based on distributional information to be consistent with world knowledge. For example, if vases are more likely to break in the real world than rocks are, we might expect *vase* to co-occur with *broke* more frequently than *rock* does. In order to test this alternative explanation, we elicited predictions for each item using a language model, BERT (Devlin, Chang, Lee, & Toutanova, 2018). We used an uncased pre-trained 'BERT for masked LM' model from the python transformers library (Wolf et al., 2020). Following Kocijan, Cretu, Camburu, Yordanov, and Lukasiewicz (2019), we elicit probabilities for each candidate by masking the pronoun and using the LM to predict the correct candidate. Where a candidate comprises multiple words, we use a corresponding number of mask tokens and find the mean of the log probabilities of each token. We normalise probabilities to the restricted decision space to obtain a measure of how biased the LM is toward NP2: *p(NP2) / (p(NP1) + p(NP2))*

Two linear mixed effects models were constructed. A base model predicted participant responses using structural bias and LM predictions. The full model additionally used physics bias to predict responses. Both models included random effects for physics and structural biases by participant. A Likelihood Ratio Test comparing the models found a significant improvement in fit from including physics bias data ($\chi^2(1) = 65.5$, $p < 0.001$). The marginal R^2 was 0.002 for the structure + LM model, and 0.56 for the full model. The full model showed positive effects of BERT predictions ($\beta = 0.604, p = 0.024$) and physics bias ($\beta = 4.58, p < 0.001$), and no effect of structural bias ($\beta = -0.313, p = 0.0487$). To ensure that the LM predictions were not adversely affected by the surface form of the experimental stimuli, we also elicited LM predictions for the physics norming stimuli (e.g. *If a vase falls on a rock, [MASK] is more likely to break*).[3] The LM normalized p(NP2) had a positive relationship with proportion of NP2 responses ($z(1) = 2.079, p = 0.037$). However, the model predicted little variance in physics norming responses (marginal R^2 = 0.014).

The result shows that physics bias explains additional variance in responses which is not accounted for by structural or distributional information. This implies an influence of world knowledge on interpretation that cannot be explained away by the alternative distributional knowledge account.

[3] We thank an anonymous reviewer for this suggestion.

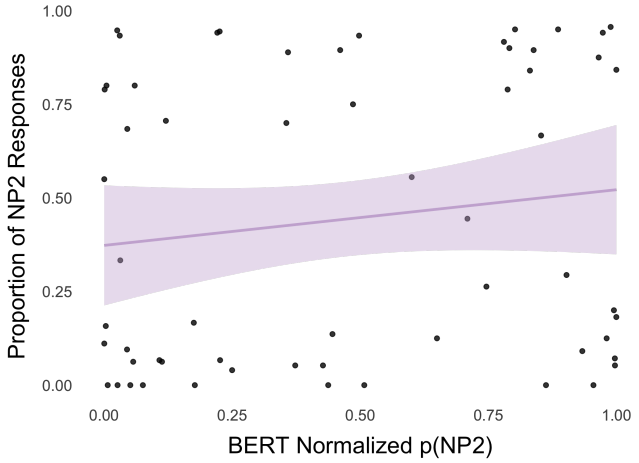

Figure 4: Language Model predictions are positively correlated with participant judgements ($r = 0.14$). However, physics bias explains additional variance which is not accounted for by LM predictions, $\chi^2(1) = 65.5$, $p < 0.001$.

Discussion

Physics bias, as measured through the physics norming study, was found to have a strong predictive effect on participant pronoun resolution decisions. Specifically, participants were more likely to select a candidate as the antecedent of a pronoun if the candidate was judged to be a more plausible participant in the described situation. In contrast, the structural bias of the sentence—exerted by grammatical features and measured in the structural norming study—did not show a significant effect on pronoun resolution decisions. Moreover, physics bias was found to improve model fit when controlling for both structural factors and distributional semantic information learned by a language model.

The results suggest that world knowledge does exert an influence on pronoun resolution. This evidence is inconsistent with the predictions of accounts which assign a minimal role to world knowledge, or a role only after a propositional interpretation for the sentence has been constructed. However, the results are consistent with accounts which incorporate world knowledge during validation or driving expectations about upcoming input, as these accounts provide a mechanism for world knowledge information to influence which referent is selected.

Implications and Future Work

The results have implications for three broad fields within cognitive science: psycholinguistics, natural language understanding, and cognitive linguistics. First, the results provide evidence against the *Minimalist* and *Construction-Integration* models in favor of the *RI-Val* and *Scenario-Mapping* models, but future work is needed to adjudicate between the latter two. One promising approach is to vary the strength of world knowledge bias. Stimuli in the present study were designed to evoke a strong world knowledge bias response (see Figure 1), however more subtle biases could be evoked by using candidates that are more similar in relevant characteristics (e.g. *When the glass fell on the vase*). Although structural factors had no effect in the present study, there is substantial evidence that they play an important role in pronoun resolution in other contexts (Crawley et al., 1990; Smyth, 1994). The **validation** account predicts that structural biases will govern resolution decisions so long as the structurally preferred candidate is not *so* implausible as to be rejected. Alternatively, an **expectation** account predicts that world knowledge will be routinely accessed and used to direct parsing, so even smaller world knowledge biases will influence pronoun resolution decisions.

Second, the results suggest understanding natural language requires general world knowledge (such as intuitions about physical properties and interactions). Much of this information is unlikely to be explicitly reported due to it being perceptually obvious and is therefore less likely to appear in text corpora used to train LMs (Shwartz & Choi, 2020). This prediction is borne out by the low proportion of variance in participant judgements explained by the LM in our results, compared to that explained by grounded information about physical biases. One solution involves developing methods to infer physical world knowledge from implicit information in corpora: for instance, the physical relationships implied by verb roles (Forbes & Choi, 2017). Alternatively, models may need to be augmented with multimodal data or datasets of physical norms (Lynott, Connell, Brysbaert, Brand, & Carney, 2019).

Finally, the results provide evidence that non-linguistic information and reasoning abilities exert influence on a core language comprehension process: reference assignment. What mechanisms and resources underlie the rapid deployment of general world knowledge during language comprehension? Battaglia, Hamrick, and Tenenbaum (2013) propose that humans are equipped with an Intuitive Physics Engine (IPE), which they can use to simulate hypothetical situations and predict their outcomes. Previous research has tested this claim on non-linguistic stimuli, but future work should examine whether the IPE can also explain physical inferences during language comprehension. Similarly, Barsalou (1999) proposes that language comprehension involves relating linguistic information to multimodal perceptual symbols grounded in sensorimotor experience. Activation of embodied perceptual symbols provides an intuitively plausible hypothesis about how world knowledge can be leveraged so efficiently to influence pronoun resolution decisions (Zwaan, 2016). Future work could explore whether sensorimotor activation correlates with evidence of world knowledge influence during language comprehension.

Acknowledgements

We thank Oisín Parkinson-Coombs, Sean Trott, James Michaelov, Tyler Chang, and Seana Coulson for insightful discussions, and four anonymous reviewers for thoughtful comments.

References

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660.

Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2007). The lme4 package. *R package version*, *2*(1), 74.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013, November). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Bender, D. (2015). Establishing a Human Baseline for the Winograd Schema Challenge. In *MAICS*.

Chambers, C. G., & Smyth, R. (1998, November). Structural Parallelism and Discourse Coherence: A Test of Centering Theory. *Journal of Memory and Language*, *39*(4), 593–608.

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

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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ferreira, F., & Yang, Z. (2019). The problem of comprehension in psycholinguistics. *Discourse Processes*, *56*(7), 485–495.

Forbes, M., & Choi, Y. (2017, July). Verb Physics: Relative Physical Knowledge of Actions and Objects. *arXiv:1706.03799 [cs]*.

Garnham, A. (2001). *Mental Models and the Interpretation of Anaphora*. Psychology Press.

Garrod, S., Freudenthal, D., & Boyle, E. (1994). The Role of Different Types of Anaphor in the On-Line Resolution of. *Journal of memory and language*, *33*, 39–68.

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic inquiry*, *5*(3), 459–464.

Gordon, P. C., & Scearce, K. A. (1995). Pronominalization and discourse coherence, discourse structure and pronoun interpretation. *Memory & Cognition*, *23*(3), 313–323.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of Word Meaning and World Knowledge in Language Comprehension. , *304*, 5.

Hartshorne, J. K. (2014). What is implicit causality? *Language, Cognition and Neuroscience*, *29*(7), 804–824.

Hobbs, J. R. (1979). Coherence and Coreference*. *Cognitive Science*, *3*(1), 67–90.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological review*, *95*(2), 163.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological review*, *85*(5), 363.

Kocijan, V., Cretu, A.-M., Camburu, O.-M., Yordanov, Y., & Lukasiewicz, T. (2019). A Surprisingly Robust Trick for Winograd Schema Challenge. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4837–4842. (arXiv: 1905.06290)

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2019). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 1–21.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological review*, *99*(3), 440.

McKoon, G., & Ratcliff, R. (2015). Cognitive theories in discourse-processing research. *Inferences during reading*, *2*.

Milburn, E., Warren, T., & Dickey, M. W. (2016). World knowledge affects prediction as quickly as selectional restrictions: evidence from the visual world paradigm. *Language, cognition and neuroscience*, *31*(4), 536–548.

O'Brien, E. J., & Cook, A. E. (2016). Coherence threshold and the continuity of processing: The RI-Val model of comprehension. *Discourse Processes*, *53*(5-6), 326–338.

Pickering, M. J., & Majid, A. (2007). What are implicit causality and consequentiality? *Language and Cognitive Processes*, *22*(5), 780–788.

Sanford, A. J., & Garrod, S. C. (1998). The role of scenario mapping in text comprehension. *Discourse processes*, *26*(2-3), 159–190.

Shwartz, V., & Choi, Y. (2020). Do Neural Language Models Overcome Reporting Bias? In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6863–6870).

Smyth, R. (1994). Grammatical determinants of ambiguous pronoun resolution. *Journal of Psycholinguistic Research*, *23*(3), 197–229.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Van den Hoven, E., & Ferstl, E. C. (2018). Discourse context modulates the effect of implicit causality on rementions. *Language and Cognition*, *10*(4), 561–594.

Willits, J. A., Amato, M. S., & MacDonald, M. C. (2015). Language knowledge and event knowledge in language use. *Cognitive psychology*, *78*, 1–27.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., ... Shleifer, S. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45).

Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review*, *23*, 1028–1034.