ProSPer: Probing Human and Neural Network Language Model Understanding of Spatial Perspective

Tessa Masis
University of Massachusetts, Amherst
140 Governors Drive
Amherst, MA 01003, USA
tmasis@umass.edu

Carolyn Jane Anderson
Wellesley College
106 Central St
Wellesley, MA 02481, USA
carolyn.anderson@wellesley.edu

Abstract

Understanding perspectival language is important for applications like dialogue systems and human-robot interaction. We propose a probe task that explores how well language models understand spatial perspective. We present a dataset for evaluating perspective inference in English, ProSPer, and use it to explore how humans and Transformer-based language models infer perspective. Although the best bidirectional model performs similarly to humans, they display different strengths: humans outperform neural networks in conversational contexts, while RoBERTa excels at written genres.

1 Introduction

Point-of-view, or **perspective**, affects many aspects of language. This paper presents ProSPer, a dataset for probing how humans and neural language models track spatial perspective in text.

In recent years, neural network language understanding has been probed in a variety of syntactic and semantic tasks.[1] We propose a probe task for one of the most complex aspects of language: relative spatial language, or **spatial perspective**. We measure the ability to infer spatial perspective using a *come*/*go* prediction task: infer a missing motion verb from a passage of text (Figure 1).

This task combines aspects of previous probe tasks (long-distance dependencies, co-reference resolution), but also poses new challenges: (1) ranking the importance of individuals in a discourse, (2) reasoning over ambiguity, and (3) inferring spatial relations. This makes it challenging for any language user: as our behavioral data shows, human performance is not perfect. However, the task may be particularly hard for language models since they lack access to grounded information, which has been hypothesized to be important for spatial language acquisition (Glenberg and Gallese, 2012).

Rick changed the subject. "I heard that you were having some furniture delivered this afternoon," he said to Aunt Emily. "I thought I'd ___ by and see if you needed any help."

(1) go **(2) come**

Figure 1: Example PRoSPer item (target: *come*)

This paper explores human and neural network language model understanding of perspective. In Sections 2-3, we motivate and present our task and dataset. In Section 4, we measure human performance on ProSPer and find accuracy rates of 77-88%. In Section 5, we evaluate pre-trained neural language models and show that the BERT family (Devlin et al., 2019) achieves human-like accuracy.

In Section 6, we explore differences between human and model behavior. Drawing on psycholinguistic work on perspective (Harris, 2012), we outline three perspective inference strategies. Our evidence supports a **genre frequency bias**: humans perform best in conversation-like contexts, while RoBERTa excels in written genres, reflecting the language each encounters most during learning.

This paper contributes to the understanding of both neural network and human language capabilities. From a cognitive science perspective, our findings contribute to two open debates: the role of grounded information in language acquisition (Section 5) and the existence of cognitive biases in perspective inference (Section 6). From an applied perspective, our results motivate greater use of conversational data in applications where perspectival language is important, such as in navigation, story generation, and human-robot interaction.

[1] Including subject-verb agreement (Linzen et al., 2016; Giulianelli et al., 2018; Gulordava et al., 2018; Lin et al., 2019), question formation (Jumelet et al., 2019; McCoy et al., 2020), filler-gap dependencies (Wilcox et al., 2018), anaphora (Jumelet et al., 2019), category membership (Ettinger, 2020), and negative polarity items (Jumelet and Hupkes, 2018).

Key contributions:

- ProSPer: a novel dataset for probing understanding of spatial perspectival language.
- Novel human behavioral data showing that humans achieve around 77-88% accuracy.
- Comparison of neural language models, showing that RoBERTa's accuracy is human-like.
- Fine-grained error analysis guided by previous psycholinguistic work, revealing a genre frequency bias for humans and RoBERTa.

2 Probing perspective inference

The perspectival motion verb prediction task is a compelling probe task because it is simple in design, but high in linguistic complexity.

2.1 Probe task description

The motion verbs *come* and *go* are identical in meaning except for point-of-view: *come* requires a discourse-important[2] person known as the **perspective-holder** to be at the destination of motion, while *go* does not. In Sentence (1) of Figure 2, for instance, motion to Boston is described with *come* because it is the location of the speaker, who is discourse-important, but motion to the beach is described with *go*.

Because *come* and *go* are synonymous except for perspective, they are an ideal test of perspective understanding for language models. If a model predicts *go* when the target is *come*, the model has failed to access an available perspective, or failed to realize the perspective-holder is at the destination. If the model predicts *come* instead of *go*, it has incorrectly inferred that the perspective-holder is at the destination.

2.2 Linguistic complexity

The perspective inference task incorporates aspects of previous probe tasks, including **long-distance dependencies** (Linzen et al., 2016), **named-entity recognition**, and **co-reference resolution** (Jumelet et al., 2019). However, perspective inference presents three additional challenges.

The first is **deciding who is important enough to be a perspective-holder**. Perspective-holders must be discourse-important, a property affected by many factors including topicality, subjecthood and definiteness (Kaiser and Lee, 2017; Hinterwimmer, 2017; Kaiser, 2020; Meuser et al., 2020). To infer perspective, therefore, a model must **gather and evaluate contextual evidence** about characters.

[2]The term **discourse prominence** is used in linguistics (Grosz et al., 1995; von Heusinger and Schumacher, 2019).

1. *Context: Sue is chatting with her sister Gina.* Sue: I'd love to see you if the flight isn't too much, but both are good options: if you **come** to Boston, you'll get a white Christmas, but if you stay in LA, you can **go** to the beach.
2. *Context: Poirot is in his flat, recounting a call from Chief Inspector Japp at Scotland Yard.* Poirot: Chief Inspector Japp thinks that the murderer will **come** to confess.

Figure 2: Constructed *come* and *go* examples

Determining discourse-importance is especially challenging because it **changes dynamically**: at any one point in a sentence, there is a unique perspective-holder, but the perspective can shift as the discourse develops, even within a sentence.

Second, the perspectival motion verb task is challenging because there can be **surface ambiguity** about the perspective-holder, as in (2) in Figure 2, where the destination could be Poirot's flat or Scotland Yard, depending on whether the perspective-holder is Poirot or Japp. Unlike syntactic tasks like subject-verb agreement, our task involves ambiguous contexts where there is no one right answer.

Third, the perspectival motion verb task involves **inferring spatial relations**. To predict *come* and *go*, a model has to both identify the perspective-holder and figure out if they are at the destination of motion. This requires inferring the motion path from the text. In (1) in Figure 2, for instance, the model has to infer that Sue lives in Boston in order to guess that the first missing verb is *come*.

2.3 Related work

Spatial language understanding is a key topic for agent-based applications. Symbolic approaches generally restrict the set of terms and perspectives considered (Winograd, 1971; Zelle and Mooney, 1996; Cangelosi et al., 2007; Boteanu et al., 2017). Statistical approaches address naturalistic language (Misra et al., 2017), but still focus on contexts with a limited set of perspectives, such as navigation (Chen et al., 2019; Paz-Argaman and Tsarfaty, 2019; Platonov et al., 2019; Zheng et al., 2020).

By contrast, the psycholinguistic community is increasingly interested in narrative contexts where the set of available perspectives may be much larger (Harris, 2012; Meuser et al., 2020). These contexts are important for computational approaches to dis-

course like story generation (Jorge et al., 2019), character tracking (Rao et al., 2015; Toshniwal et al., 2020), and stance detection (Augenstein et al., 2016; Inkpen et al., 2017).

Spatial language is also a topic of importance to language acquisition. The extent to which grounded information is necessary for language acquisition is much debated within the cognitive science (Rehm et al., 2003; Glenberg and Gallese, 2012) and language modeling communities (Lucy and Gauthier, 2017; Bender and Koller, 2020; Bisk et al., 2020). Evidence from child language acquisition shows that learners rely on non-linguistic situational cues (Clark, 1973; Samuelson et al., 2011), but cannot establish if such information is necessary. If it is, spatial perspective is one of the most likely phenomena to require it, since it is relative and situational. Our proposed task explores this issue: successful perspective inference by text-based neural networks would imply that grounded information is not necessary for language acquisition.

Our task therefore brings together questions from several research communities. Using methods drawn from the neural network probe task and psycholinguistics communities, we gather novel empirical evidence that addresses two debated issues in cognitive science: grounding in language acquisition (Section 5), and cognitive biases in perspective shift (Section 6). In addition, our findings are relevant to several application areas, including work on embodied agents and story generation.

3 ProSPer: Probing Spatial Perspective

We present a new corpus for probing human and neural network ability to infer spatial perspective: ProSPer.[3]. The ProSPer dataset consists of two parts: the Automatic subset, a large set of examples extracted by string-matching, and the Annotated subset, a smaller set for fine-grained analysis. The Annotated items were hand-selected for linguistic diversity and annotated with linguistic features.

3.1 Task

PRoSPer probes perspective inference with a forced-choice task: given a passage with an omitted verb, decide if the missing word is *come* or *go*. Figures 1 and 3 show examples. In addition to the critical perspectival verbs, *come* and *go*, we also include three non-perspectival comparison verbs: *walk*, *drive*, and *arrive*. Each is compared with its closest semantic competitor: *come* and *go* with each other, *walk* and *drive*, and *arrive* with *come*.

3.2 Automatically selected subset

The Automatic subset consists of 47385 examples taken from the Open American National Corpus using all forms of *come*, *go*, *walk*, *arrive*, and *drive*.

3.3 Annotated subset

The Annotated subset consists of 600 examples from publicly available corpora of American English.[4] Examples were chosen to avoid non-perspectival uses (like *Come on, man!*) and to include a variety of genres. The examples were annotated by the authors with the following features:

Perspective-holder: Examples containing *come* were annotated by perspective-holder category: speaker, listener, attitude-holder, protagonist, theme, empathy center, home-base, or accompaniment. A quota system was used to ensure that each category was well-represented.

Subject: The perspective-holder is often ambiguous in sentences with non-perspectival verbs. Instead, the subject of the verb was recorded. These should not be confused: the perspective-holder is rarely the subject, since *come* requires the perspective-holder to be located at the destination of motion (Barlew, 2017).

Syntactic environment: Examples were categorized based on the verb's syntactic environment: top-level, within the scope of a speech verb, within the scope of a thought verb, quotation, or other. This is important because attitude verbs like *say* and *think* increase the importance of their subjects' perspectives. A quota system was used to ensure that each environment was represented.

Destination of motion: The destination of motion was recorded for all examples.

Tense: Examples in both subsets of ProSPer were sorted into coarse tense categories.

4 Human performance

We measured human performance on a representative subset of ProSPer using a forced choice task similar to bidirectional language modeling.[5]

[3]The ProSPer dataset and code can be found in the ProSPer Github repository. Files relating to the human experiments can be found in the ProSPer Open Science Foundation repository.

[4]The Corpus of Contemporary American English, the Corpus of Online Registers of English, and The Corpus of American Soap Operas (Davies, 2008, 2016, 2011).

[5]The IRB-approved design was preregistered through the Open Science Foundation. Experimental stimuli, results, and

Human format: *and they worked with her and got her walking and got her taking kind of taking care of herself and so she was able to ___ back home*
Bidirectional format: *and they worked with her and got her walking and got her taking kind of taking care of herself and so she was able to **MASK** back home*
Unidirectional format: *and they worked with her and got her walking and got her taking kind of taking care of herself and so she was able to*

Figure 3: ProSPer presentation formats, target: *go*

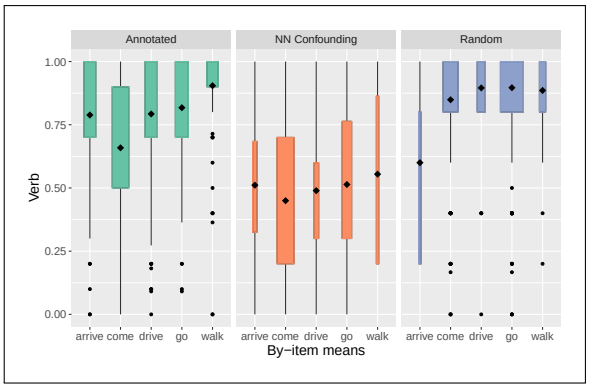

Figure 4: Human item means by verb type and corpus (diamond=mean, box=1st-3rd quartiles)

4.1 Experimental design

Task: Participants were shown an item with the target omitted and asked to pick between two verbs (Figure 3). Each target verb was presented with its closest semantic competitor (see Section 3).
Items: Human judgments were collected on three subsets of ProSPer: the entire Annotated subset; 600 items from the Automatic subset, sampled randomly (Random); and the 300 most challenging items from the Automatic subset (NN Confounding).[6]
Participants: 300[7] monolingual American English speakers were recruited on Prolific. Each participant saw 20 Annotated items, 10 NN Confounding items, and 10 Random items, as well as 30 filler items. Participants were randomly assigned to item lists using a Latin square design. Each Annotated and NN Confounding item was seen by 10 participants, while Random items were seen by 5 participants.

4.2 Results

Mean accuracies by corpus and verb are shown in Figure 4. Participants did best on the Random subset (88.1% accuracy) and worst on the NN Confounding subset (47.2%). However, mean accuracies varied considerably by verb type, with *come* generally proving harder than other verbs.

5 Neural language model performance

We evaluated the performance of various neural network language models on the ProSPer dataset.

5.1 Baselines

We provide two baseline models trained on English Wikipedia: a Kneser-Ney smoothed trigram model and a unidirectional LSTM (Hochreiter and Schmidhuber, 1997).[8]

5.2 Transformer-based models

We compare the performance of pre-trained Transformer-based language models from the HuggingFace Transformers library (Wolf et al., 2019): Transformer-XL (Dai et al., 2019), BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), GPT (Radford, 2018), and GPT-2 (Radford et al., 2018). Since our goal is to evaluate how much knowledge of spatial perspective models acquire from the language modeling task, we do not do fine-tune any of the models.

It is important to note that the different families of models are not directly comparable, since the BERT family of models are bidirectional and condition their predictions on a larger context. The models also differ in complexity and training data; retraining each model on the same dataset would be prohibitive in terms of computational cost.[9]

5.3 Task

The models' performance on ProSPer was measured in a two-way comparison, where the relative probability of the target word was compared to that of its closest semantic competitor. For unidirectional language models, we take the probabilities of each word at the target site. For bidirectional

analysis scripts can also be found there.

[6]We pooled the 1000 hardest items for each Transformer model and then selected the items missed by the most models.

[7]Excluding participants who did not meet the language criteria, had less than 80% accuracy on the attention check fillers, or gave an incoherent response to a bot check question.

[8]Implementations adapted from Rescia (2015) and Verwimp et al. (2018).

[9]Further description of the models can be found in the Appendix. Although the training data varies by model, we have verified that there is no overlap with ProSPer.

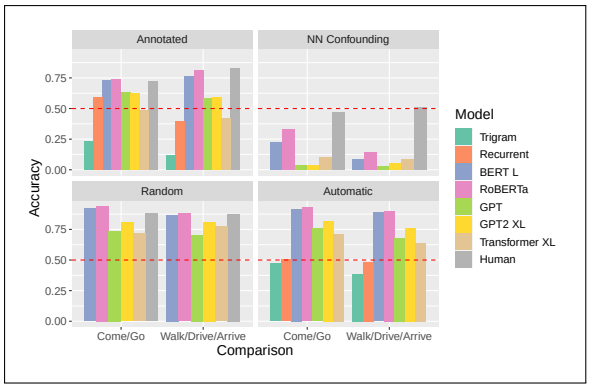

Figure 5: NN means by corpus[10]

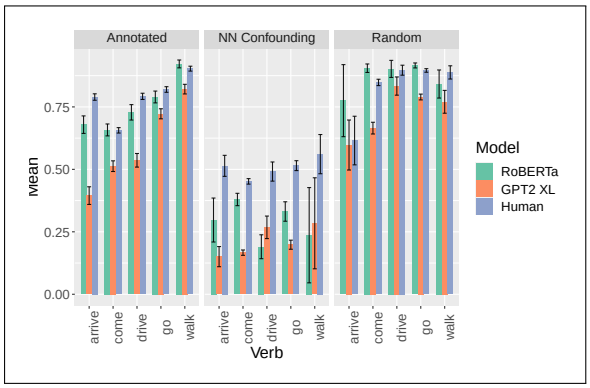

Figure 6: Human and NN means by corpus and verb

models, the target word was masked and the entire context was presented (Figure 3).

5.4 Results

Model performance on each subset of ProSPer is shown in Figure 5. The results suggest that neural language models are capable of tracking and using spatial perspective to some degree. Although all models find the Annotated subset more challenging than the Automatic one, the best-performing model, RoBERTa, performs similarly to human participants on the Annotated subset, and slightly better on the Random Automatic subset.

The GPT and GPT-2 models perform well above chance on the Automatic dataset but struggle on the harder Annotated dataset, suggesting that their success on the Automatic dataset may be built on non-perspectival uses of *come* and *go*. However, it is important to note that human performance was measured on the bidirectional task.

While the baseline models perform poorly,

[10]The below-chance NN Confounding results may be unintuitive, but since this subset was selected by poor model performance, it is expected (selection is causally dependent on low score).

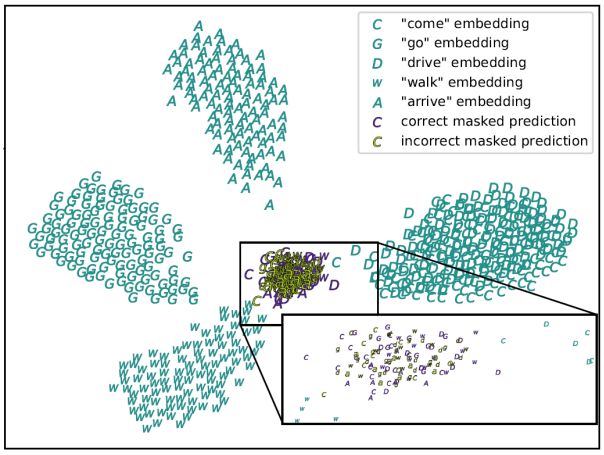

Figure 7: PCA of RoBERTa Automatic verb and Annotated MASK hidden layers. Inset: MASK cluster.

BERT models perform similarly to humans. This is impressive given the task's linguistic complexity: it suggests that models are able to use context clues to track the spatial perspectives and discourse-importance of characters. However, these results should not be over-interpreted: although we tried to control for possible confounds with the Annotated dataset, we cannot rule out the possibility that the models relying on some kind of simpler heuristic.[11]

5.5 Principal Components Analysis

To explore how neural networks represent the key motion verbs, we provide a principal components analysis of RoBERTa's hidden states retrieved after substituting each of the five target verbs into examples from the Automatic dataset.

Figure 7 shows the first two principal components for RoBERTa's hidden states when given each verb and when given masked examples from the Annotated dataset (*Susan is <mask> here*). We see that *come* and *go* are on opposing sides of the plot, reflecting their relationship as a minimal pair.

The masked tokens are clustered in the center of the plot and relatively equidistant from each verb cluster. We note that masked tokens from examples that RoBERTa classifies correctly have a slightly broader spread than incorrect ones. This may reflect the model's relative uncertainty.

6 How do human and neural network errors differ?

As Figure 6 shows, the best neural network model (RoBERTa) performs similarly to humans in the

[11]Bender and Koller (2020) discuss some cautionary tales.

Strong Egocentricity Hypothesis

- Low accuracy for *come* relative to other verbs.

Weak Egocentricity Hypothesis

- High accuracy with speaker perspectives.

Genre Frequency Bias Hypothesis

- Human accuracy improved by conversation-like contexts (spoken genres, quoted environments) and speaker or listener perspectives.
- RoBERTa accuracy improved by text-like contexts and attitude holder, empathy center, theme, and protagonist perspectives.

Figure 8: Perspective Inference Strategy Predictions

Random and Annotated subsets.[12] Although human participants found the NN Confounding subset difficult, with performance at about chance, they outperformed all models (unsurprising, since the subset was selected based on average model difficulty).

Averaging across neural network models, there is a medium correlation between by-item model and human performance (ρ=0.65); RoBERTa's predictions correlate a little more closely with human performance (ρ=0.71). The medium correlation suggests that despite their similar accuracy, there is considerable variance between human and RoBERTa predictions on ProSPer.

In the remainder of the paper, we explore possible explanations for these differences.

6.1 Three perspective inference strategies

To gain insight into how humans and neural network models infer perspective, we explore three possible perspective inference strategies.

Previous psycholinguistic work on perspective suggests that humans are **egocentric**: they are biased toward their own perspectives. The core idea is that accessing a self perspective is automatic, while accessing other perspectives involves a perspective shift operation (Epley et al., 2004; Lin et al., 2010). This causes slower, more errorful processing of non-self-oriented perspectival language.

Although there is converging evidence for egocentric cognitive bias from a variety of tasks,[13] there is ongoing debate over its strength (Brown-Schmidt and Heller, 2018). A strong version of the egocentricity hypothesis posits that participants are so self-biased that they struggle to access any other perspective, no matter how discourse-important. This **Strong Egocentricity Hypothesis** predicts low accuracy for human participants on all ProSPer *come* items.

Other psycholinguists propose an indirect egocentricity effect. If listeners are aware of speaker egocentricity, they may proactively take the speaker's perspective to facilitate processing (Harris, 2012; Anderson, 2020). We refer to this as the **Weak Egocentricity Hypothesis**: participants assume the perspective holder is the speaker because they know speakers tend to be egocentric. This hypothesis makes different predictions about human behavior on ProSPER. Rather than performing poorly on *come* in general, we expect accuracy to vary based on the perspective-holder: participants should predict *come* accurately when the speaker is the perspective holder, but underperform when non-speaker perspectives are discourse-important.

We introduce a third possibility: the **Genre Frequency Bias**. We note that although humans and RoBERTa both encounter a variety of data during language acquisition, its composition differs: children learn through conversation, while RoBERTa is primarily trained on news, non-fiction and third-person narratives. The only portion of RoBERTa's training data that contains conversation-like text is the Open Web Text corpus, and it differs in significant ways from in-person conversation (i.e., speakers lack mutual awareness of each other's spatial location). We hypothesize that this leads to advantages on different subsets of ProSPer.

Since humans acquire language through conversation, they may predict *come* best when anchored to the perspective of a conversation participant: a speaker or listener. They may also excel at transcribed speech. Conversely, RoBERTa may perform better in written genres and with third-person perspectives like protagonists and discourse themes. We also expect differences by syntactic environment: quotation is like conversation, while speech and belief contexts are more common in text.

The Annotated subset of ProSPer makes it easy to test the predictions of these three hypotheses (Figure 8). Note that they are not mutually exclusive: for instance, we could find evidence of both an egocentricity bias and a genre frequency bias

[12] To compare to human performance, the models' predicted probabilities for the two verb choices for each item have been renormalized to produce a prediction score from 0 to 1.

[13] Including two-player reference tasks (Horton and Keysar, 1996; Hanna et al., 2003; Heller et al., 2008), eye-tracking tasks (Brown-Schmidt and Hanna, 2011; Ferguson et al., 2017; Child et al., 2020), self-paced reading tasks (Millis, 1995), and interpretation tasks (Harris, 2012; Köder et al., 2015).

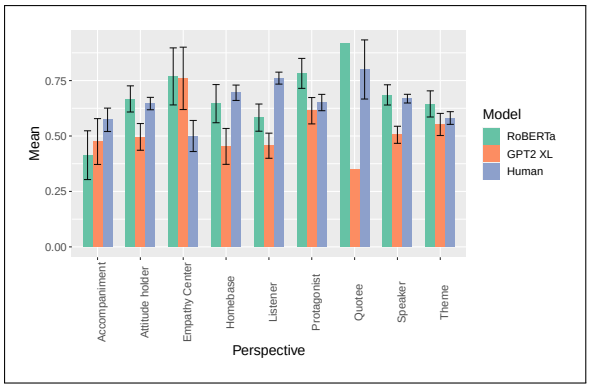

Figure 9: Annotated *come* means by perspective type

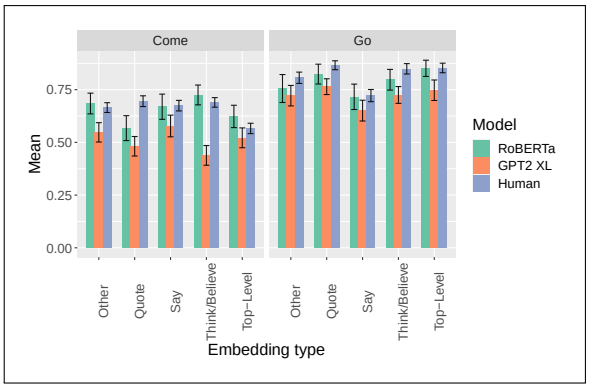

Figure 11: Annotated means by syntactic environment

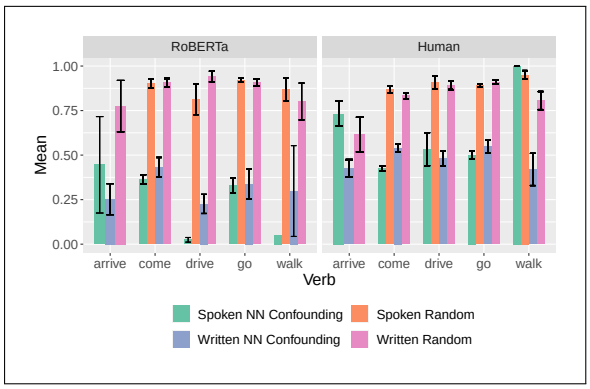

Figure 10: Means by modality and corpus

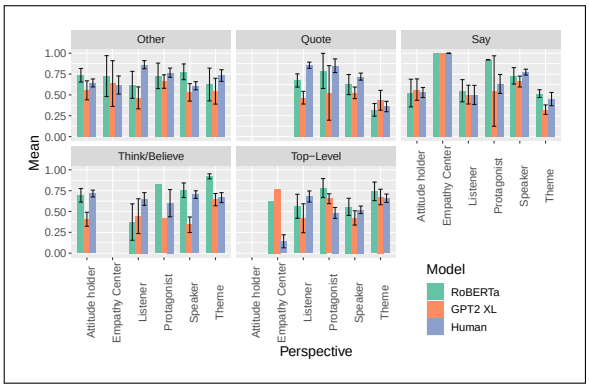

Figure 12: Annotated *come* means by perspective and syntactic environment

(lower *come* accuracy relative to other verbs, but mediated by perspective holder and genre). We explore the evidence for each, discussing only trends that are supported by statistically significant effects in regression models.[14]

6.2 Strong Egocentricity Evidence

The Strong Egocentricity Hypothesis predicts low accuracy for *come* items overall, due to interference from participants' own perspectives.

We do find that human performance on *come* is significantly lower than the verb group mean. However, we also find significantly lower accuracy on *come* for RoBERTa. It is unclear how the Strong Egocentricity Hypothesis could explain this, since neural networks do not hold a self-perspective. Our evidence is inconclusive: there may be a human Strong Egocentricity bias and another reason for RoBERTa's performance, or *come* items may challenge both groups for other reasons, such as the perspective ambiguity they introduce.

[14]Details of the regression models and full results are in the Appendix. Figure bars show standard error.

6.3 Weak Egocentricity Evidence

The Weak Egocentricity Hypothesis predicts that performance on *come* items will be highest when the perspective holder is the speaker. This prediction is not borne out by the human results. Although there were significant negative effects of empathy center, theme, and protagonist perspectives on human *come* performance, there was no significant positive effect of speaker perspective (Figure 9). Interestingly, there was an overall positive effect of speaker on RoBERTa's performance.

6.4 Genre Frequency Bias Evidence

The Genre Frequency Bias Hypothesis makes several predictions. First, it predicts that humans should excel in conversational contexts. We compared performance on the spoken and written subsets of the Random dataset. For humans, there was a significant positive effect of the spoken subset for *come* (Figure 10). The reverse was true for RoBERTa: accuracy at predicting *come* was signif-

icantly worse in the spoken subset.[15]

Second, the Genre Frequency Bias Hypothesis predicts that humans will perform best when the perspective holder is a speaker or listener, while RoBERTa will perform better with empathy center, attitude holder, theme, and protagonist perspectives. Our results support these predictions to some extent. Humans did significantly worse with empathy center, them, and protagonist perspectives, but RoBERTa did not do significantly better with them. Humans also performed better with listener perspectives than attitude holder perspectives, but there was no significant different between human performance with speaker and attitude holder perspectives. In addition, as noted above, there was a positive effect of speaker perspective for RoBERTa.

The Genre Frequency Bias Hypothesis also extends to syntactic environments: humans are expected to perform best in conversation-like contexts, such as quotation and top-level clauses. We observed that quoted environments had an overall positive effect on human performance, but found a negative effect of top-level contexts on *come* accuracy (Figure 11). However, this effect is complicated by the proportional representation of perspective holders in the Annotated subset. In a model including both perspective holder type and syntactic environment, a negative interaction with top-level environments is found only for empathy center, them, and protagonist perspectives (Figure 12).

RoBERTa's performance on *come* was improved by belief environments, and hurt by top-level and quoted environments. Mirroring the human data, we find a positive interaction between empathy center, theme, and protagonist perspectives and top-level and quoted environments for RoBERTa.

Thus, our data partially supports the Genre Frequency Bias. For humans, performance is better in conversation-like contexts and worse with third-person perspectives. For RoBERTa, performance is worse in conversation-like contexts. However, there was no preference for speaker perspectives over attitude holders for humans, as predicted, and RoBERTa performed well with speaker perspectives, which is unexpected if exposure to conversational contexts is what leads to speaker bias.

[15]This effect was significant in a RoBERTa-only model, but not in a human-RoBERTa model where RoBERTa was treated as a fixed effect. The human effects were significant in both.

6.5 Manual error analysis

To further explore possible perspective inference strategies, we examined *come*/*go* errors in the 10 hardest items for humans in the Annotated and Random subsets. Most of the Random errors involve incomplete sentences. The remaining 7 *come*/*go* examples in the Random and Annotated subsets involve errors consistent with the Genre Frequency Bias. In four, the target is *come*, but the speaker is the subject. This may lead speaker-biased participants astray: if the speaker is the perspective-holder, *go* must be used. One error involves *go* as a target for a context where the speaker is at the destination, which again, contradicts expectations if the speaker is the perspective-holder. The last two involve third-person perspective-holders.

7 Conclusion

In this paper, we present ProSPer, a dataset for probing how humans and language models infer spatial perspective from text. We use ProSPer's *come*/*go* prediction task to gather novel psycholinguistic data and evaluate Transformer-based language models. We find that the best bidirectional model, RoBERTa, performs similarly to humans, providing tentative evidence that grounded information is not required for spatial language acquisition.

Despite near-equal accuracy, we find only a medium correlation between RoBERTa and human scores, suggesting different perspective inference strategies. We explored two strategies proposed in prior psycholinguistic work, but ultimately argued in support of a novel Genre Frequency Bias effect: humans perform best in conversation-like contexts, while RoBERTa performs best in third-person narrative contexts. Although many of the observed trends support the Genre Frequency Bias, it did not explain the full pattern of results. We hope that ProSPer's Annotated subset will aid future work exploring perspective inference strategies.

The observed Genre Frequency Bias is important for both neural network and human language understanding. For cognitive scientists, it fuels an existing debate on cognitive biases in perspective inference. For applications where human-like use of perspectival language is important, like navigation, story generation, and human-robot interaction, it suggests that conversational training data may improve model robustness.

References

Carolyn Jane Anderson. 2020. *Shifting the perspectival landscape*. Ph.D. thesis, University of Massachusetts, Amherst.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *EMNLP*.

Jefferson Barlew. 2017. *The semantics and pragmatics of perspectival expressions in English and Bulu: The case of deictic motion verbs*. Ph.D. thesis, The Ohio State University.

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.

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

Adrian Boteanu, Jacob Arkin, Siddharth Patki, Thomas M. Howard, and Hadas Kress-Gazit. 2017. Robot-initiated specification repair through grounded language interaction. *CoRR*, abs/1710.01417.

Sarah Brown-Schmidt and Joy E. Hanna. 2011. Talking in Another Person's Shoes: Incremental Perspective-taking in Language Processing. *Dialogue and Discourse*, 2:11–33.

Sarah Brown-Schmidt and Daphna Heller. 2018. Perspective-taking During Conversation. In G Gaskell and S. A. Rueschemeyer, editors, *Oxford Handbook of Psycholinguistics*. Oxford University Press.

A. Cangelosi, V. Tikhanoff, J. F. Fontanari, and Emmanouil Hourdakis. 2007. Integrating language and cognition: A cognitive robotics approach. *IEEE Computational Intelligence Magazine*, 2:65–70.

Howard Chen, Alane Suhr, Dipendra Kumar Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12530–12539.

Scarlett Child, Jane Oakhill, and Alan Garnham. 2020. Tracking your Emotions: an Eye-Tracking Study on Reader's Engagement with Perspective during Text Comprehension. *Quarterly Journal of Experimental Psychology*.

Eve V. Clark. 1973. Non-linguistic strategies and the acquisition of word meanings. *Cognition*, 2(2):161 – 182.

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.

Mark. Davies. 2008. *The Corpus of Contemporary American English (COCA): 600 million words, 1990-present*. Available online at https://www.english-corpora.org/coca.

Mark. Davies. 2011. *Corpus of American Soap Operas: 100 million words*. Available online at https://www.english-corpora.org/soap.

Mark. Davies. 2016. *Corpus of Online Registers of English (CORE)*. Available online at https://www.english-corpora.org/core.

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

Nicholas Epley, Boaz Keysar, Leaf Van Boven, and Thomas Gilovich. 2004. Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, 87(3):327–339.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Heather J. Ferguson, Ian Apperly, and James E. Cane. 2017. Eye tracking reveals the cost of switching between self and other perspectives in a visual perspective-taking task. *The Quarterly Journal of Experimental Psychology*, 70(8):1646–1660.

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.

Arthur. M. Glenberg and Vittorio Gallese. 2012. Action-based language: A theory of language acquisition, comprehension, and production. *Cortex*, 48.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: a framework for modeling the local coherence of discourse. *Association for Computational Linguistics*, pages 203–225.

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.

Joy E. Hanna, Michael K. Tanenhaus, and John C. Trueswell. 2003. The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, 49(1):43–61.

Jesse A. Harris. 2012. *Processing Perspectives*. Ph.D. thesis, University of Massachusetts, Amherst.

Daphna Heller, Daniel Grodner, and Michael K. Tanenhaus. 2008. The role of perspective in identifying domains of reference. *Cognition*, 108(3):831–836.

Stefan Hinterwimmer. 2017. Prominent protagonists. *Journal of Pragmatics*.

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

William S. Horton and Boaz Keysar. 1996. When do speakers take into account common ground? *Cognition*, 59(1):91–117.

Diana Inkpen, Xiao-Dan Zhu, and Parinaz Sobhani. 2017. A dataset for multi-target stance detection. In *EACL*.

A. Jorge, Ricardo Campos, A. Jatowt, and S. Bhatia. 2019. The 2nd international workshop on narrative extraction from text: Text2story 2019. In *ECIR*.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics.

Jaap Jumelet, Willem Zuidema, and Dieuwke Hupkes. 2019. Analysing neural language models: Contextual decomposition reveals default reasoning in number and gender assignment. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1–11, Hong Kong, China. Association for Computational Linguistics.

Elsi Kaiser. 2020. Shifty behavior: Investigating predicates of personal taste and perspectival anaphora. In *Semantics and Linguistic Theory*, volume 30.

Elsi Kaiser and Jamie Herron Lee. 2017. Experience matters: A psycholinguistic investigation of predicates of personal taste. *Semantics and Linguistic Theory*, 27.

Franziska Köder, Emar Maier, and Petra Hendriks. 2015. Perspective shift increases processing effort of pronouns: a comparison between direct and indirect speech. *Language, Cognition and Neuroscience*, 30(8).

Shuhong Lin, Boaz Keysar, and Nicholas Epley. 2010. Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, 46:551–556.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Li Lucy and Jon Gauthier. 2017. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 76–85, Vancouver, Canada. Association for Computational Linguistics.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

Sara Meuser, Stefan Hinterwimmer, and Maximilian Hörl. 2020. Online Processing of Protagonists' Perspective-Taking. In *The CUNY Sentence Processing Conference*, volume 33.

Keith K. Millis. 1995. Encoding discourse perspective during the reading of a literary text. *Poetics*, 23(3):235–253.

Dipendra Kumar Misra, John Langford, and Yoav Artzi. 2017. Mapping instructions and visual observations to actions with reinforcement learning. *CoRR*, abs/1704.08795.

Tzuf Paz-Argaman and Reut Tsarfaty. 2019. RUN through the streets: A new dataset and baseline models for realistic urban navigation. 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 6449–6455, Hong Kong, China. Association for Computational Linguistics.

Georgiy Platonov, Benjamin Kane, Aaron Gindi, and Lenhart K. Schubert. 2019. A spoken dialogue system for spatial question answering in a physical blocks world. *CoRR*, abs/1911.02524.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Sudha Rao, Allyson Ettinger, Hal Daumé III, and Philip Resnik. 2015. Dialogue focus tracking for zero pronoun resolution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 494–503, Denver, Colorado. Association for Computational Linguistics.

M. Rehm, K. Rohlfing, and K. U. Goecke. 2003. Situatedness: The interplay between context(s) and situation. *Journal of Cognition and Culture*, 3:132–156.

Giovanni Rescia. 2015. Procesamiento de lenguaje natural.

L.K. Samuelson, L.B. Smith, L.K. Perry, and J.P. Spencer. 2011. Grounding word learning in space. *PLoS ONE*, 6.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Shubham Toshniwal, Allyson Ettinger, Kevin Gimpel, and Karen Livescu. 2020. PeTra: A Sparsely Supervised Memory Model for People Tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5415–5428, Online. Association for Computational Linguistics.

Lyan Verwimp, Hugo Van hamme, and Patrick Wambacq. 2018. Tf-lm: Tensorflow-based language modeling toolkit. *Proceedings of LREC*.

Klaus von Heusinger and Petra B. Schumacher. 2019. Discourse prominence: Definition and application. *Journal of Pragmatics*, 154:117 – 127.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Terry Winograd. 1971. *Procedures as a representation of data in a computer program for understanding natural language*. PhD thesis, Massachusets Institute of Technology.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

John M. Zelle and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, AAAI'96, page 1050–1055. AAAI Press.

Kaiyu Zheng, Deniz Bayazit, Rebecca Mathew, Ellie Pavlick, and Stefanie Tellex. 2020. Spatial language understanding for object search in partially observed cityscale environments.

A ProSPer composition

A.1 Automatic subset

All instances of each verb that occurred in the OANC and MASC corpora were included in the Automatic subset.

	Modality	OANC	MASC	Total
come	Spoken	5222	142	12650
	Written	6854	432	
go	Spoken	19337	427	28841
	Written	8471	606	
walk	Spoken	812	12	2098
	Written	1100	174	
arrive	Spoken	23	0	843
	Written	770	50	
drive	Spoken	1182	6	2953
	Written	1681	84	
Total		45452	1933	47385

Table 1: Summary of Automatic subset by source and modality

Source	*come*	*go*	*walk*	*arrive*	*drive*	Total
court transcript	40	115	3	0	1	159
debate transcript	56	136	5	0	3	200
face-to-face	591	1552	119	11	111	2384
telephone	4677	17961	697	12	1073	24420
blog	28	70	4	0	16	118
email	12	22	1	2	5	42
essays	22	25	0	0	9	56
ficlets	56	80	43	5	4	188
fiction	123	134	66	3	21	347
government	280	366	22	37	123	828
jokes	50	83	21	5	9	168
journal	4780	6739	440	338	921	13218
letters	171	26	18	3	8	226
movie script	29	55	42	10	0	136
newspaper	25	33	2	7	6	73
non-fiction	333	200	23	22	91	669
spam	23	26	0	2	0	51
technical	579	500	52	169	279	1579
travel guides	827	575	538	215	265	2490
twitter	25	66	2	2	8	103
Total	12650	28841	2098	843	2953	47385

Table 2: Summary of Automatic subset examples by genre

A.2 Annotated subset

The Annotated examples were selected using a quota system. A minimum of 15 examples were selected in each syntactic environment (25 for 'come'). A minimum of 20 examples of 'come' were selected for each main perspective-holder type, balanced across syntactic environments.

B ProSPer annotation schema

B.1 Perspective-holder categories for Annotated *come* items

Speaker: current speaker is the perspective-holder.

	None	*say*	*believe*	Quote	Other	Total
Speaker@ET	4	4	4	4	4	20
Speaker@UT	4	4	4	4	4	20
Protagonist	4	1	1	2	2	12
Listener	5	3	3	7	3	21
Attitude-holder	0	8	8	0	7	21
Home-base	1	1	1	1	1	5
Accompaniment	1	1	1	1	1	5
Other	6	3	3	6	3	21
Total	25	25	25	25	25	125

Table 3: Summary of *come* examples in Annotated subset by syntactic environment and perspective-holder

	None	*say*	*believe*	Quote	Other	Total
Speaker	4	3	3	4	3	17
Protagonist	3	0	0	1	1	5
Listener	3	3	3	5	2	16
Attitude-holder	0	5	5	0	4	14
3rd-person	3	2	2	3	3	13
Home-base	1	1	1	1	1	5
Accompaniment	1	1	1	1	1	5
Total	15	15	15	15	15	75

Table 4: Summary of non-*come* examples in Annotated subset by syntactic environment and subject

Example: *The sunlight falls into the room through the top windows. It's a big classroom, holds one fifty in a marble building. The oldest on campus. The pillars and doorways chiseled Grecian columns and archways. I would **come** to class even if I didn't like learning the stories and culture of southern Africa just to look at the sun glancing off the marble and dashing off the swaying bouncing leaves outside the windows. It's hot in here.*

Listener: the listener of the current speaker is the perspective-holder.

Example:

Attitude-holder: the subject of an attitude verb is the perspective-holder.

Example:

Protagonist: the main character of a story is the perspective-holder.

Example: *Walking home from Sudbury one night, he became aware of galloping horses, then he saw lights **coming** toward him and he stepped off the road to let the carriage pass . He saw distinctly one man or two in the box, driving; they had no heads, only hands and the lower part of the body. As he watched, the whole thing vanished.*

Theme: the person whose is the topic of the discourse is the perspective-holder.

Example: *Dante: Well, thank you for your patience, but we got to find the other Karen Anderson. It's pretty urgent. Karen: Oh. I hope for her sake there's no problem with her son. Dante: Actually, her son is in trouble, and we have reason to believe that he'll **come** to her for help.*

Empathy center: a person who the speaker cares strongly about (usually a family member) is the perspective-holder.

Example: *"We need your help." "She hasn't opened the blind in three days," he said. I raised my voice in case he couldn't hear me over the hurdy-gurdy music and the dull throb of the crowd. "Valerie is in labor." But he continued to stare at the window, black marble eyes filled with something like pain and hope, joy and terror and want. I imagined that death was the color of those eyes; I have often thought it since, but I yelled at him to **come** anyway, that Valerie was having a baby, that we needed him, that the baby was more important than this.*

Home-base: the motion is towards the home-base of the perspective-holder (i.e, house or workplace).

Example: *as i am new to this sport and am currently living in edinburgh it is easy for me to find some good routes to ride but i am **coming** home to nottingham soon for a couple of weeks and am very keen to ride whilst i am home. Could someone please suggest some routes around nottingham (if there are any) for me to ride my email is REDACTED.*

Accompaniment: the motion is alongside the perspective-holder.

Example: *Madagascar was one of those films that I really wanted to go and see at the cinema but nobody would **come** with me and my sister had already been to see it so I had to wait for the video.*

B.2 Syntactic environment categories for Annotated items

Top-level: verb appears in the matrix clause.

Example: *On Saturday I **went** with a friend and her husband to the Guggenheim Museum in Bilbao, a Basque city on the northern edge of Spain that has been revitalized by the stunning titanium and marble building by Frank Gehry. To call it a building does not do it justice. This is no ordinary museum. It is alive, a structure in motion, with titanium scales that move as the light washes over them.*

Speech verb: verb is in the scope of a speech verb like *say* or *tell*.

Example: *Mimi. I love your work. Found you through. Purl Bee. I visit your web site for inspiration and to "visit". I'm in California but daughters boyfriend is from Boston I've said if I ever **go** there I want to see your work or take a workshop.*

Thought verb: verb is in the scope of a thought verb like *think* or *believe*.

Example: *Stephanie: You saved a seat for me. Thank you. I thought I'd find you here. Nick: I'm not going, you know. They asked me to go, but I'm not going. Stephanie: You don't think I **drove** all the way down here to try and talk you into going, do you? Nick: If you came here for some Brooke bashing, you came to the wrong place. Let me tell you that right now. Stephanie: Actually, I came for a drink.*

Quotation: verb is part of a quotation.

Example: *Although he hasn't put all the pieces together yet, young Justin knows that words have power, too. "Once we stopped at a paint store with my mother," Laurie Bradley recalls. "She **went** in to get something while we waited in the car. Justin noticed the closed sign in the door of the store that the owner had forgotten to flip over. He said, "Mommy that's a c. That means kids aren't allowed. Don't you just hate that?""*

Other: verb is in the scope of a predicate that is not a speech or thought verb, such as *discover* or *learn*.

Example: *I see the mailman's truck go by and I know his routine so well; he'll be at my house in fifteen minutes. "I got ta go," I say, and standing, I take of my Thriftway apron. "I wanted to ask you to **go** to the movies," he says and I don't have time to think. I have fourteen minutes to check out with my boss and run home.*

B.3 Tense/aspect categories for Annotated and Automatic subsets

1: infinitive, 1st/2nd person habitual
come, go, walk, drive, arrive

2: progressive
coming, going, walking, driving, arriving

3: simple past
came, went, walked, drove, arrived

4: 3rd person habitual
comes, goes, walks, drives, arrives

5: past present
come, gone, walked, drove, arrived

C Human study

C.1 Data selection

Human performance was measured on the entire Annotated subset and two subsets of Automatic subset: the Random and NN Confounding subsets.

The Random subset was randomly sampled from the full Automatic subset; a handful of examples were excluded and resampled because they were deemed offensive or upsetting.

The NN Confounding subset consists of examples that proved challenging for neural network models. To select this subset, the items in the Automatic subset were ranked for each model by difficulty, and the top 100 most difficult items for each model were collated. These items were then re-ranked by how many lists they appeared on (how many models found them very difficult), and the top 100 were selected.

C.2 Participant instructions

The instructions to participants were as follows:

> You will see a passage of text where one of the words is missing (shown as a blank).
>
> You will be given several options for words that could fill in the blank. **Please select the word that you think is the best fit in the blank.** Don't worry about giving a perfect answer— just go with your first impression of the sentence.
>
> You can either click on the word to select it or use the number keys on your keyboard to select it: if you think the first word is the best fit, hit the "1" button on your keyboard.
>
> Most of the time, there will be only one word missing, but some times there will be two blanks. When that happens, you will have to choose a pair of words to fill in the blanks.
>
> **Note**: some sentences were automatically collected from online sources, including audio transcripts. **Because of this, you may notice spelling errors, incomplete sentences, speech hesitations ('uh', 'um') or weird punctuation.** Please try to ignore this.
>
> Here are a couple of practice items before the main experiment begins.

C.3 Practice items

Participants were given a chance to practice on the following items. They were not given feedback; this was just to acclimate them to the task format and response keys. The expected responses are in bold.

- One of the students in the class hadn't turned in his homework. Ms. Morris frowned and thought about what to do. After school, she called his mother to ___ her.
  (a) **tell** (b) fail

- I asked Harriet how the book was getting on and if Peter's suggestions had helped. Helen said, "Oh, yes, you write, don't you?" as if she'd never heard of her, and asked what the title was, so that she could get it from the library. Harriet said, quite gravely, "That is very kind of you, but do let me ___ you one– I am allowed six free copies, you know." First sign of temper, but I don't blame her.
  (a) **send** (b) buy
- "How about some breakfast?" asked Rupert, jerking a thumb back towards the tent. He fetched a box of wood chips from the back of the van and in a surprisingly short order had an admirable campfire crackling away in the churchyard. Next, he produced a coffeepot, a packet of bacon, a loaf of bread, and a couple of sharpened sticks to make a toast with. Nialla had even managed to find a jar of ___ somewhere in their baggage.
  (a) **marmalade** (b) almond butter

C.4 Data exclusion criteria

The data of participants who did not meet the language criteria, had less than 80% accuracy on the attention check fillers, or gave an incoherent response to a bot check question was excluded (the participants were, of course, still compensated). The attention check and bot check questions are given below (correct responses in bold).

C.4.1 Attention check items

- Davis handed Jasmine an large square box wrapped in festive paper and topped with an enormous pink bow. When she unwrapped it and opened the lid, she gasped. Inside was a beautiful crimson dress. The silk rustled as she lifted it carefully out of the box and held it up to the light. "It's perfect!" she declared. "How did you know that ___ is my favorite color?"
  (a) **red** (b) pink
- Polly's cat was a very finicky eater. At the pet store, she couldn't find the particular flavor of cat food that Percy liked: Seafood Surprise. After stopping by the grocery store, she got lunch ready: a tuna sandwich for ___ and a bowl of plain tuna for ___.
  (a) Percy/Polly (b) **Polly/Percy**
- Benjamin's favorite thing about where he lived was the desert. He could drive in it for hours, stopping his truck wherever he felt like it, and wait for night to fall. Lying in his truck bed, Benjamin could watch the ___ for hours, feeling utterly at peace.
  (a) fireworks (b) **stars**
- Betty had been working secretly on a beautiful patchwork quilt for months in her studio. When she gave it her mother at Christmas, Maia was amazed by its bright swirling colors. Since it had taken so long to make, Betty was very pleased that her mother liked the ___.
  (a) **blanket** (b) painting
- Throughout her writing, Anne had a unique talent for pushing literary boundaries and blending genres. She also tended to incorporate whatever strange new topic she had recently taken an interest in (the last one was Marilyn Monroe). Whatever she published, her readers knew to expect something ___.
  (a) **eccentric** (b) traditional
- Margaret celebrated her 80th birthday last month with unadulterated gaiety. During the party, she was often heard saying that the only thing she missed about being ___ was street hockey.
  (a) **young** (b) old
- Mr. Dougal hated catching cheaters, and Billy hated taking quizzes. This week, Billy was glad that he was sitting next to Sally, the best student in the class. After noticing how closely their answers matched, Mr. Dougal gave Billy an ___.
  (a) **F** (b) A
- Our mother, Harriet, to whom Buckshaw had been left by her uncle Tarquin de Luce, had died in a mountaineering accident in the Himalayas when I was a year old. Because she had left no ___, the vultures of His Majesty's Board of Inland Revenue had descended upon Father at once, and had been busily pecking out his liver ever since.
  (a) **will** (b) children
- Sheryl loved her son Clayton, but sometimes his forgetfulness drove her crazy. Thursdays were always extra busy at the bakery. Thursdays were also basketball nights, and Clayton had forgotten his uniform: again. Sheryl sighed. During her lunch break, ___ managed to sneak away and drop it off in the main office.
  (a) **she** (b) he
- Samantha was looking forward to her birthday party. It was going to be at the Cincinnati Zoo. Her dad was describing what kinds of animals there would be. He thought there ___ be any lions there. "That's ok. What I really want to see is a penguin."
  (a) would (b) **wouldn't**
- "So where are you taking me for brunch?" her mom asked. Abigail picked up her purse from the couch and said, "I was thinking we'd go to Green Bean. The French toast there is amazing!" "You remember that I'm allergic to nuts, right?" said her mom skeptically. "Yeah, but I don't think there'll be ___ in the French toast!"
  (a) some (b) **any**

C.4.2 Bot check questions

The bot check questions were designed to elicit free text responses from participants. Any participants who answered incoherently were excluded. We recognize that this is a subjective judgment. However, in practice, these participants almost always had low accuracy on the attention check items as well. Moreover, since we paid participants regardless of whether their data was ultimately included, we felt that it was acceptable to be overly strict.

- Imagine that you are back in elementary school. You've just made a new friend and invited him to come over to your house after school. How would you give him directions from your school to your house? (Please don't share any actual addresses or other identifying information; we're just interested in how people give directions.)
- What is your favorite insult and why?

We also excluded participants who gave incoherent answers to the demographic questions that were asked (for instance, "USA" in response to "What state do you currently live in?" or "None" in response to "How old are you?").

D Language model details

Model complexity and training data details are shown in Tables 5 and 6.

Family	Model	Layers	Attention heads	Embedding size
Transformer-XL	base	18	16	1024
BERT	base	12	12	768
	large	24	16	1024
RoBERTa	base	12	12	768
DistilBERT	base	6	12	768
GPT	base	12	12	768
GPT-2	base	12	12	768
	medium	24	16	1024
	large	36	20	1280
	extra-large	48	25	1600

Table 5: Model complexity

Family	Training data	# tokens (M)	Vocabulary
Transformer-XL	WikiText-103	100	26735
BERT	English Wikipedia, BooksCorpus	3300	30522
RoBERTa	English Wikipedia, BooksCorpus CC-News, Open Web Text, Stories	$>$ 3300	50266
DistilBERT	English Wikipedia, BooksCorpus	3300	30522
GPT	BooksCorpus	800	40478
GPT-2	WebText	unknown	50257

Table 6: Model training data

E Detailed results

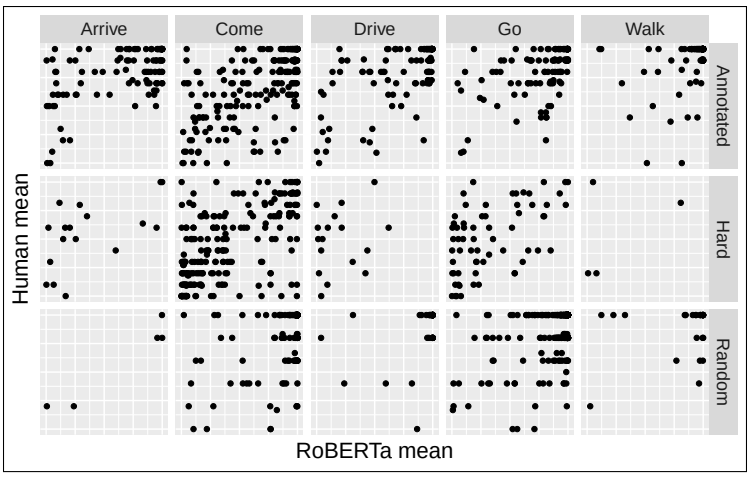

Figure 13: Human item means versus RoBERTa scores

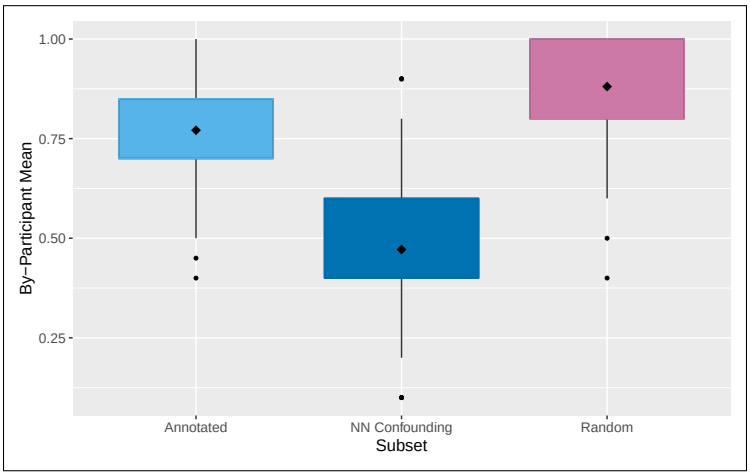

Figure 14: Participant means by subset, diamond=mean, interval=1st-3rd quartile

Condition	Overall mean	*come/go* mean	*walk/drive/ arrive* mean
Annotated	77.1%	73.8%	82.8%
Random	88.1%	87.2%	80.0%
NN Confounding	47.2%	48.3%	52.2%

Table 7: Human performance on ProSPer

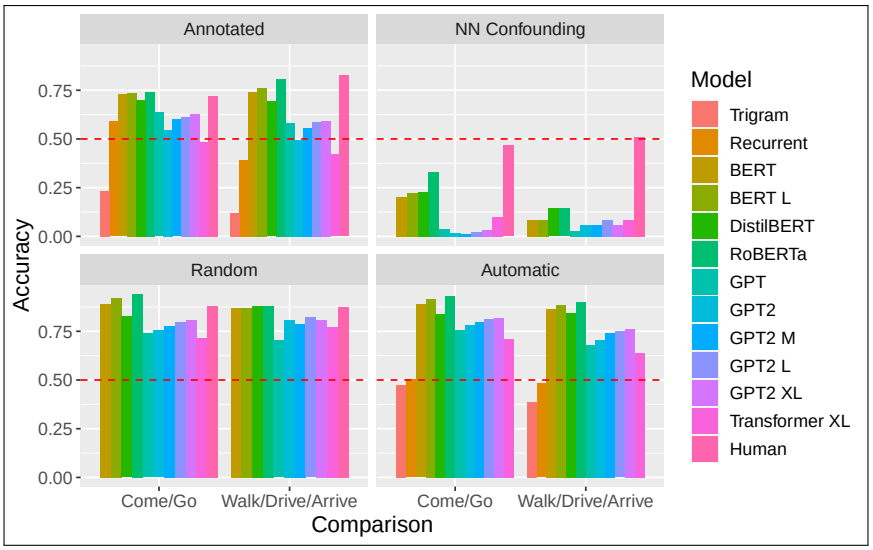

Figure 15: NN means by corpus

Family	Model	Corpus	Accuracy	*come/go* accuracy	*walk/drive/arrive* accuracy
random guess	-	-	20%	50%	50%
trigram	forward	Annotated	12.2%	23.3%	11.9%
		Automatic	40.7%	47.3%	38.6%
	backward	Annotated	27.5%	53.8%	27.0%
		Automatic	40.7%	46.9%	33.6%
RNN	wiki	Annotated	29.5%	59.0%	39.2%
		Automatic	21.6%	50.3%	48.2%

Table 8: Language model performance on ProSPer

Family	Model	Corpus	Accuracy	*come*/*go* accuracy	*walk*/*drive*/*arrive* accuracy
Transformer-XL	base	Annotated	28.0%	48.5%	42.0%
		Automatic	63.4%	71.2%	64.0%
		Random	-	71.7%	77.4%
		NN Confounding	-	9.8%	8.6%
BERT	base	Annotated	55.7%	72.8%	74.1%
		Automatic	85.0%	88.7%	86.3%
		Random	-	89.2%	86.9%
		NN Confounding	-	20.3%	8.6%
	large	Annotated	58.3%	73.4%	76.1%
		Automatic	88.2%	91.5%	88.7%
		Random	-	92.2%	86.9%
		NN Confounding	-	22.2%	8.6%
RoBERTa	base	Annotated	**63.5**%	**74.1**%	**80.9**%
		Automatic	**89.7**%	**93.0**%	**90.0**%
		Random	-	**93.9**%	**88.1**%
		NN Confounding	-	**33.2**%	**14.3**%
DistilBERT	base	Annotated	53.0%	69.8%	69.3%
		Automatic	78.7%	83.7%	84.6%
		Random	-	82.6%	**88.1**%
		NN Confounding	-	22.6%	**14.3**%
GPT	base	Annotated	39.8%	63.6%	58.0%
		Automatic	68.5%	75.6%	68.0%
		Random	-	73.8%	70.2%
		NN Confounding	-	3.8%	2.9%
GPT-2	base	Annotated	32.8%	54.4%	49.5%
		Automatic	71.7%	78.1%	70.6%
		Random	-	75.4%	81.0%
		NN Confounding	-	1.9%	5.7%
	medium	Annotated	38.5%	60.0%	55.6%
		Automatic	73.9%	80.0%	74.1%
		Random	-	77.9%	78.6%
		NN Confounding	-	1.1%	5.7%
	large	Annotated	40.2%	61.3%	58.7%
		Automatic	75.2%	81.1%	74.9%
		Random	-	79.8%	82.1%
		NN Confounding	-	2.3%	8.6%
	extra-large	Annotated	**41.0**%	**62.6**%	**59.4**%
		Automatic	**75.9**%	**81.8**%	**76.2**%
		Random	-	**80.6**%	**81.0**%
		NN Confounding	-	**3.4**%	**5.7**%

Table 9: Language model performance on ProSPer

F Detailed PCA results

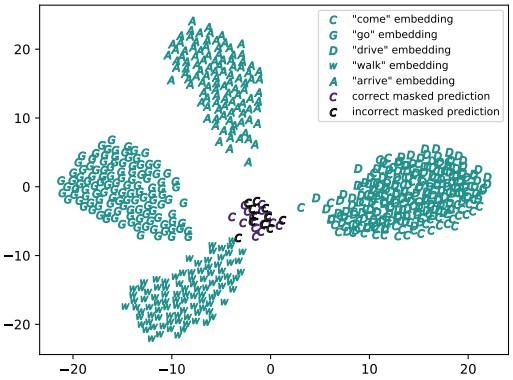

Figure 16: RoBERTa trained on Automatic subset and tested on masked "come" embeddings from Annotated corpus

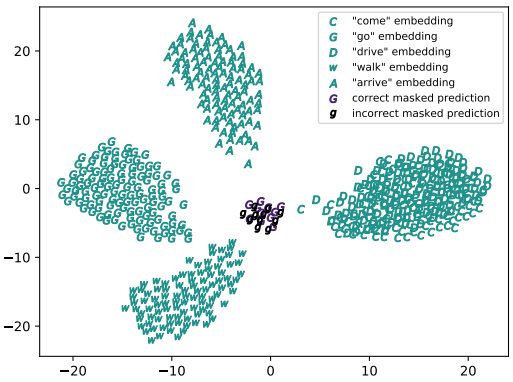

Figure 17: RoBERTa trained on Automatic subset and tested on masked "go" embeddings from Annotated corpus

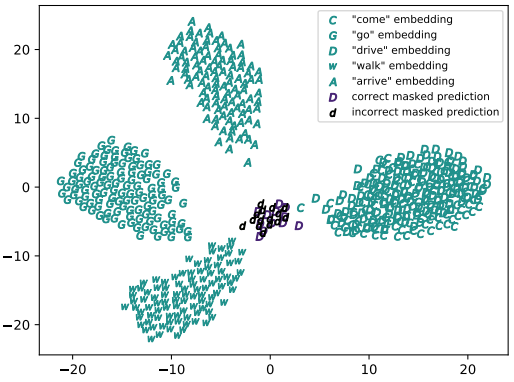

Figure 18: RoBERTa trained on Automatic subset and tested on masked "drive" embeddings from Annotated corpus

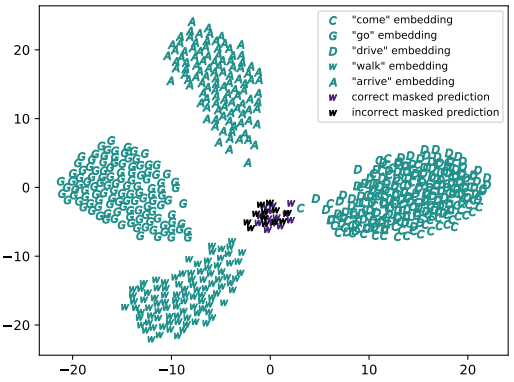

Figure 19: RoBERTa trained on Automatic subset and tested on masked "walk" embeddings from Annotated corpus

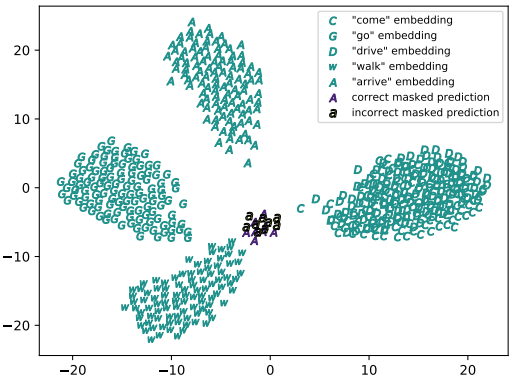

Figure 20: RoBERTa trained on Automatic subset and tested on masked "arrive" embeddings from Annotated corpus

G Regression results

G.1 Human models

Predictors of human scores were explored in a series of mixed-effects logistic regression models. Because our items were not controlled by condition, we do not include by-item random effects.

Fixed effects (n=12160)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.2	0.041	30.2	5.997e-200
come	-0.97	0.094	-10.4	3.2e-25
go	-0.51	0.1	-5.01	5.312e-07
drive	-0.57	0.1	-5.58	2.47e-08
arrive	-0.64	0.1	-6.3	3.071e-10
c.annotated.corpus	-0.59	0.075	-7.81	5.931e-15
n.annotated.corpus	-0.42	0.064	-6.56	5.502e-11
n.hard	-1.4	0.068	-20.9	1.177e-96
c.hard	-1.2	0.07	-17	6.038e-65

Table 10: Human performance by corpus and verb, corpus nested in *come*

Fixed effects (n=12160)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.2	0.054	22.5	3.324e-112
come	-1.2	0.15	-8.06	7.777e-16
go	-0.44	0.16	-2.8	0.005185
drive	-0.6	0.17	-3.46	0.0005442
arrive	-0.96	0.19	-5.11	3.271e-07
c.annotated.corpus	0.0031	0.055	0.0561	0.9553
c.spoken.corpus	-0.46	0.054	-8.64	5.677e-18
n.spoken.corpus	0.82	0.16	5.02	5.273e-07
n.annotated.corpus	0.65	0.076	8.6	8.02e-18
go:n.spoken.corpus	-1.4	0.3	-4.59	4.379e-06
go:n.annotated.corpus	-0.85	0.15	-5.75	9.016e-09
drive:n.spoken.corpus	-0.97	0.34	-2.9	0.003784
drive:n.annotated.corpus	-0.56	0.16	-3.63	0.0002838
arrive:n.spoken.corpus	-0.56	0.37	-1.51	0.1316
arrive:n.annotated.corpus	0.011	0.17	0.0621	0.9505

Table 11: Human performance by modality and verb, modality nested in *come* and Written treated as baseline

Fixed effects (n=6080)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.5	0.042	35.5	2.196e-276
come	-1.3	0.12	-10.5	7.954e-26
attitude.b	-0.059	0.15	-0.383	0.7016
listener.b	0.44	0.18	2.48	0.01325
anchor.b	-0.062	0.16	-0.389	0.6972
other.b	-0.32	0.13	-2.47	0.01354
go	-0.62	0.12	-5.3	1.141e-07
drive	-0.72	0.12	-6.21	5.212e-10
arrive	-0.78	0.11	-6.95	3.668e-12

Table 12: Human performance by perspective holder, Speaker treated as baseline

Fixed effects (n=12160)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.7	18	0.0937	0.9254
c.hard.spoken.corpus	-1.3	0.091	-13.8	2.276e-43
c.hard.written.corpus	-0.88	0.11	-8.07	7.115e-16
c.random.spoken.corpus	0.55	0.16	3.5	0.0004685
c.annotated.mod.corpus	-0.44	0.09	-4.9	9.687e-07
n.hard.spoken.corpus	3.5	1.4e+02	0.0242	0.9807
n.hard.written.corpus	-1.5	0.18	-7.9	2.812e-15
n.random.spoken.corpus	0.72	0.31	2.34	0.01953
n.annotated.mod.corpus	0.12	0.14	0.87	0.3845
come	-2.7	72	-0.0377	0.97
go	-2.3	72	-0.0321	0.9744
drive	-2.3	72	-0.0316	0.9748
arrive	-2.5	72	-0.035	0.972
n.hard.spoken.corpus:go	-9.7	2.9e+02	-0.0338	0.973
n.hard.spoken.corpus:drive	-9.5	2.9e+02	-0.0332	0.9735
n.hard.spoken.corpus:arrive	-7.8	2.9e+02	-0.0271	0.9784
n.hard.written.corpus:go	-0.16	0.34	-0.467	0.6403
n.hard.written.corpus:drive	-0.29	0.37	-0.768	0.4426
n.hard.written.corpus:arrive	0.66	0.43	1.54	0.1239
n.random.spoken.corpus:go	-1.1	0.41	-2.66	0.007775
n.random.spoken.corpus:drive	-0.8	0.51	-1.55	0.1202
n.annotated.mod.corpus:go	-1	0.24	-4.27	1.968e-05
n.annotated.mod.corpus:drive	-1.1	0.28	-3.79	0.0001536
n.annotated.mod.corpus:arrive	0.023	0.35	0.0653	0.9479

Table 13: Human performance by corpus, modality and verb, corpus+modality nested in *come* and Written Random treated as baseline

Fixed effects (n=6080)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.5	0.041	37	2.917e-300
come	-1.3	0.11	-12.3	5.884e-35
c.believe	0.095	0.12	0.787	0.4312
c.say	0.013	0.13	0.103	0.918
c.quote	0.11	0.13	0.823	0.4103
c.none	-0.36	0.12	-3.04	0.00239
go	-0.64	0.12	-5.21	1.85e-07
drive	-0.73	0.12	-6.03	1.665e-09
arrive	-0.77	0.12	-6.44	1.16e-10
n.believe	0.21	0.14	1.49	0.1363
n.say	0.12	0.15	0.801	0.4231
n.quote	0.28	0.15	1.88	0.06046
n.none	0.1	0.13	0.814	0.4159
go:n.believe	0.031	0.29	0.108	0.9139
go:n.say	-0.65	0.3	-2.17	0.0297
go:n.quote	-0.14	0.31	-0.434	0.664
go:n.none	0.22	0.27	0.831	0.4058
drive:n.believe	-0.33	0.28	-1.19	0.2354
drive:n.say	-0.091	0.31	-0.294	0.7687
drive:n.quote	-0.29	0.31	-0.931	0.3517
drive:n.none	0.16	0.26	0.609	0.5426
arrive:n.believe	0.32	0.3	1.06	0.2879
arrive:n.say	-0.5	0.3	-1.66	0.09612
arrive:n.quote	-0.56	0.31	-1.81	0.07074
arrive:n.none	0.045	0.26	0.174	0.8617

Table 14: Human performance by syntactic environment and verb, environment nested in *come* and Other treated as baseline

Fixed effects (n=6080)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.5	0.044	34.9	4.197e-266
come	-1.3	0.13	-10.1	5.596e-24
c.believe	0.31	0.25	1.25	0.2112
c.say	0.7	0.24	2.94	0.003318
c.quote	0.46	0.24	1.92	0.05508
c.none	-0.29	0.22	-1.31	0.1913
listener.b	0.27	0.19	1.41	0.1572
anchor.b	-0.1	0.17	-0.612	0.5405
other.b	-0.33	0.14	-2.42	0.01554
attitude.b	-0.1	0.2	-0.513	0.6077
go	-0.65	0.12	-5.23	1.654e-07
drive	-0.74	0.12	-6.06	1.325e-09
arrive	-0.78	0.12	-6.46	1.016e-10
n.believe	0.21	0.14	1.5	0.1339
n.say	0.12	0.15	0.816	0.4148
n.quote	0.28	0.15	1.86	0.06291
n.none	0.1	0.13	0.804	0.4217
c.believe:listener.b	-1.3	0.5	-2.52	0.0119
c.believe:anchor.b	0.67	0.41	1.63	0.1036
c.believe:other.b	-0.66	0.36	-1.84	0.06568
c.say:listener.b	-2.1	0.56	-3.84	0.0001251
c.say:anchor.b	0.76	0.46	1.64	0.09999
c.say:other.b	-1.4	0.36	-3.9	9.477e-05
c.quote:listener.b	-0.45	0.48	-0.935	0.3498
c.quote:anchor.b	0.46	0.46	1	0.3157
c.quote:other.b	-1.4	0.35	-3.96	7.405e-05
c.none:listener.b	-0.46	0.48	-0.958	0.3382
c.none:anchor.b	0.98	0.42	2.33	0.01977
c.none:other.b	-0.41	0.31	-1.34	0.1792
c.believe:attitude.b	0.015	0.34	0.0444	0.9646
c.say:attitude.b	-1.1	0.35	-3.22	0.001268
go:n.believe	0.024	0.29	0.085	0.9323
go:n.say	-0.65	0.3	-2.18	0.02892
go:n.quote	-0.14	0.31	-0.436	0.6629
go:n.none	0.22	0.27	0.822	0.411
drive:n.believe	-0.34	0.28	-1.21	0.2255
drive:n.say	-0.098	0.31	-0.316	0.7517
drive:n.quote	-0.29	0.31	-0.933	0.3506
drive:n.none	0.16	0.27	0.592	0.5542
arrive:n.believe	0.31	0.31	1.01	0.3102
arrive:n.say	-0.51	0.3	-1.69	0.0909
arrive:n.quote	-0.56	0.31	-1.83	0.06706
arrive:n.none	0.041	0.26	0.16	0.8727

Table 15: Human performance by perspective holder and syntactic environment, perspectives compared to group mean

Fixed effects (n=6080)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.5	0.042	36.7	2.444e-294
come	-1.3	0.12	-10.9	6.689e-28
c.believe	0.06	0.14	0.433	0.6648
c.say	-0.084	0.15	-0.544	0.5867
c.quote	0.28	0.19	1.48	0.1389
c.none	-0.076	0.17	-0.436	0.6626
listener	0.3	0.18	1.72	0.08627
speaker	0.084	0.14	0.612	0.5405
other	-0.18	0.14	-1.32	0.1875
attitude	0.29	0.21	1.4	0.1612
go	-0.65	0.12	-5.23	1.654e-07
drive	-0.74	0.12	-6.06	1.325e-09
arrive	-0.78	0.12	-6.46	1.016e-10
n.believe	0.21	0.14	1.5	0.1339
n.say	0.12	0.15	0.815	0.4148
n.quote	0.28	0.15	1.86	0.06292
n.none	0.1	0.13	0.804	0.4217
c.believe:listener	-1.5	0.43	-3.56	0.0003661
c.believe:speaker	-0.53	0.33	-1.63	0.1036
c.believe:other	-1.1	0.34	-3.16	0.001604
c.say:listener	-2.3	0.51	-4.53	5.971e-06
c.say:speaker	-0.6	0.37	-1.64	0.1
c.say:other	-1.7	0.38	-4.52	6.25e-06
c.quote:listener	-0.73	0.46	-1.6	0.1102
c.quote:speaker	-0.37	0.37	-1	0.3157
c.quote:other	-1.5	0.37	-3.97	7.175e-05
c.none:listener	-1.1	0.44	-2.62	0.008723
c.none:speaker	-0.78	0.34	-2.33	0.01977
c.none:other	-1.1	0.33	-3.41	0.0006576
c.believe:attitude	-0.52	0.32	-1.62	0.1053
c.say:attitude	-1.5	0.38	-3.98	6.748e-05
go:n.believe	0.024	0.29	0.085	0.9323
go:n.say	-0.65	0.3	-2.18	0.02892
go:n.quote	-0.14	0.31	-0.436	0.6629
go:n.none	0.22	0.27	0.822	0.411
drive:n.believe	-0.34	0.28	-1.21	0.2255
drive:n.say	-0.098	0.31	-0.316	0.7517
drive:n.quote	-0.29	0.31	-0.933	0.3507
drive:n.none	0.16	0.27	0.592	0.5542
arrive:n.believe	0.31	0.31	1.01	0.3102
arrive:n.say	-0.51	0.3	-1.69	0.0909
arrive:n.quote	-0.56	0.31	-1.83	0.06706
arrive:n.none	0.041	0.26	0.16	0.8727

Table 16: Human performance by perspective holder and syntactic environment, perspective compared to group mean

G.2 RoBERTa models

Predictors of RoBERTa scores were explored in a series of regression models.

Fixed effects (n=1495)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.64	0.02	32.3	9.901e-174
come	-0.043	0.055	-0.777	0.4375
c.annotated.corpus	-0.14	0.026	-5.23	1.938e-07
c.hard	-0.38	0.026	-14.6	2.428e-45
go	0.025	0.056	0.438	0.6614
drive	-0.085	0.061	-1.39	0.1641
arrive	-0.097	0.07	-1.38	0.1679
n.annotated.corpus	-0.0061	0.03	-0.206	0.8371
n.hard	-0.43	0.059	-7.2	9.836e-13
go:n.annotated.corpus	-0.089	0.048	-1.85	0.06428
go:n.hard	-0.0022	0.11	-0.0207	0.9835
drive:n.annotated.corpus	-0.14	0.055	-2.54	0.01134
drive:n.hard	-0.14	0.12	-1.16	0.2478
arrive:n.annotated.corpus	-0.075	0.084	-0.888	0.3749
arrive:n.hard	0.044	0.14	0.321	0.7484

Table 17: RoBERTa performance by corpus and verb, corpus nested in *come*

Fixed effects (n=1495)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.69	0.025	28.1	3.324e-139
come	-0.13	0.052	-2.52	0.01168
c.spoken.corpus	-0.15	0.03	-5.12	3.401e-07
c.annotated.mod.corpus	-0.026	0.037	-0.712	0.4768
n.spoken.corpus	0.022	0.055	0.396	0.692
n.annotated.mod.corpus	0.14	0.044	3.15	0.001693
go	0.041	0.052	0.786	0.4322
drive	-0.068	0.061	-1.11	0.2652
arrive	-0.27	0.089	-3.02	0.002565
n.spoken.corpus:go	-0.046	0.083	-0.554	0.5795
n.spoken.corpus:drive	-0.04	0.1	-0.394	0.6939
n.spoken.corpus:arrive	-0.12	0.16	-0.734	0.4628
n.annotated.mod.corpus:go	-0.13	0.084	-1.52	0.1276
n.annotated.mod.corpus:drive	-0.11	0.089	-1.2	0.2316
n.annotated.mod.corpus:arrive	0.018	0.1	0.175	0.8609

Table 18: RoBERTa performance by modality and verb, Written treated as baseline and modality nested in *come*

Fixed effects (n=1495)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.6	0.026	22.8	1.529e-98
come	0.032	0.074	0.439	0.6605
c.hard.spoken.corpus	-0.45	0.038	-11.9	3.209e-31
c.hard.written.corpus	-0.37	0.052	-7.03	3.223e-12
c.random.spoken.corpus	0.061	0.047	1.29	0.1962
c.annotated.mod.corpus	-0.14	0.037	-3.89	0.000106
n.hard.spoken.corpus	-0.57	0.13	-4.24	2.344e-05
n.hard.written.corpus	-0.47	0.085	-5.57	3e-08
n.random.spoken.corpus	0.023	0.072	0.314	0.7532
n.annotated.mod.corpus	0.0046	0.046	0.101	0.9194
go	0.068	0.074	0.913	0.3615
drive	-0.055	0.084	-0.654	0.5133
arrive	-0.028	0.096	-0.297	0.7666
n.hard.spoken.corpus:go	0.16	0.25	0.621	0.5345
n.hard.spoken.corpus:drive	-0.079	0.28	-0.277	0.7819
n.hard.spoken.corpus:arrive	0.26	0.29	0.876	0.3814
n.hard.written.corpus:go	-0.058	0.16	-0.353	0.724
n.hard.written.corpus:drive	-0.25	0.17	-1.44	0.1496
n.hard.written.corpus:arrive	-0.042	0.19	-0.219	0.8263
n.random.spoken.corpus:go	-0.0073	0.096	-0.0757	0.9396
n.random.spoken.corpus:drive	-0.084	0.12	-0.714	0.4756
n.annotated.mod.corpus:go	-0.11	0.081	-1.34	0.1817
n.annotated.mod.corpus:drive	-0.2	0.088	-2.29	0.02195
n.annotated.mod.corpus:arrive	-0.11	0.11	-0.968	0.3334

Table 19: RoBERTa performance by corpus, modality, and verb, Written Random corpus treated as baseline and corpus+modality nested in nested in *come*

Fixed effects (n=1495)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.79	0.017	45.2	1.349e-191
come	-0.24	0.042	-5.72	1.72e-08
c.believe	0.14	0.077	1.83	0.06844
c.say	0.099	0.08	1.23	0.2196
c.other_e	0.12	0.077	1.5	0.1348
c.none	0.022	0.078	0.279	0.7806
go	-0.078	0.046	-1.69	0.09241
drive	-0.16	0.047	-3.33	0.0009359
arrive	-0.2	0.048	-4.09	4.859e-05
n.believe	-0.0023	0.057	-0.0395	0.9685
n.say	-0.022	0.058	-0.374	0.7084
n.other_e	-0.011	0.058	-0.183	0.8548
n.none	0.024	0.054	0.439	0.6607
go:n.believe	0.018	0.12	0.153	0.8784
go:n.say	-0.033	0.12	-0.279	0.7806
go:n.other_e	0.00015	0.12	0.00124	0.999
go:n.none	0.072	0.11	0.646	0.5184
drive:n.believe	0.017	0.12	0.144	0.8852
drive:n.say	0.0084	0.12	0.0678	0.9459
drive:n.other_e	0.076	0.13	0.606	0.5449
drive:n.none	0.11	0.12	0.935	0.3501
arrive:n.believe	0.13	0.13	1.04	0.2995
arrive:n.say	-0.062	0.13	-0.494	0.6218
arrive:n.other_e	0.094	0.13	0.747	0.4551
arrive:n.none	0.17	0.12	1.45	0.1462

Table 20: RoBERTa performance by syntactic environment and verb, environment nested in *come*

Fixed effects (n=600)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.8	0.019	41.1	2.389e-175
come	-0.2	0.055	-3.69	0.000247
attitude.b	-0.035	0.092	-0.381	0.7033
listener.b	-0.062	0.098	-0.631	0.5285
anchor.b	-0.16	0.097	-1.69	0.0915
other.b	-0.015	0.079	-0.197	0.8441
go	-0.074	0.046	-1.63	0.1044
drive	-0.16	0.047	-3.32	0.0009536
arrive	-0.19	0.048	-4.07	5.382e-05

Table 21: RoBERTa performance by perspective-holder, Speaker treated as baseline and perspective nested in *come*

Fixed effects (n=1495)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.8	0.019	41.1	1.233e-170
come	-0.2	0.055	-3.69	0.0002461
c.believe	-1.6e-15	0.14	-1.11e-14	1
c.say	-0.011	0.13	-0.0852	0.9321
c.quote	-0.13	0.14	-0.925	0.3553
c.none	-0.24	0.14	-1.73	0.08406
listener.b	-0.12	0.11	-1.12	0.2616
anchor.b	-0.2	0.098	-2.08	0.03781
other.b	-0.017	0.081	-0.21	0.8341
attitude.b	-0.093	0.12	-0.776	0.4383
go	-0.078	0.046	-1.71	0.0884
drive	-0.16	0.047	-3.37	0.0008086
arrive	-0.2	0.048	-4.15	3.927e-05
n.believe	0.0084	0.058	0.147	0.8833
n.say	-0.011	0.058	-0.188	0.8507
n.quote	0.011	0.058	0.185	0.853
n.none	0.034	0.054	0.635	0.5259
c.believe:listener.b	-0.44	0.26	-1.69	0.09253
c.believe:anchor.b	0.2	0.25	0.814	0.4158
c.believe:other.b	0.33	0.21	1.58	0.1155
c.say:listener.b	-0.23	0.3	-0.756	0.45
c.say:anchor.b	0.25	0.26	0.985	0.3251
c.say:other.b	0.12	0.2	0.569	0.5696
c.quote:listener.b	0.11	0.23	0.493	0.6222
c.quote:anchor.b	-0.069	0.27	-0.257	0.7975
c.quote:other.b	-0.036	0.2	-0.175	0.8614
c.none:listener.b	0.08	0.25	0.323	0.7465
c.none:anchor.b	1.2e-15	0.26	4.61e-15	1
c.none:other.b	0.39	0.18	2.08	0.0376
c.believe:attitude.b	-0.086	0.2	-0.433	0.665
c.say:attitude.b	-0.29	0.21	-1.37	0.171
go:n.believe	0.018	0.12	0.148	0.8822
go:n.say	-0.033	0.12	-0.274	0.7845
go:n.quote	-0.00015	0.12	-0.00126	0.999
go:n.none	0.072	0.11	0.627	0.5306
drive:n.believe	-0.059	0.12	-0.49	0.6245
drive:n.say	-0.068	0.12	-0.545	0.5863
drive:n.quote	-0.076	0.12	-0.613	0.5398
drive:n.none	0.032	0.12	0.274	0.7842
arrive:n.believe	0.039	0.13	0.314	0.7533
arrive:n.say	-0.16	0.12	-1.27	0.2041
arrive:n.quote	-0.094	0.12	-0.757	0.4494
arrive:n.none	0.076	0.11	0.67	0.5032

Table 22: RoBERTa performance by syntactic environment and perspective-holder, Speaker treated as baseline and environment and perspective nested in *come*

Fixed effects (n=1495)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.78	0.018	43.5	2.628e-181
come	-0.28	0.046	-5.95	4.771e-09
c.believe	0.0014	0.077	0.0186	0.9852
c.say	-0.041	0.087	-0.47	0.6383
c.quote	-0.14	0.11	-1.34	0.1803
c.none	-0.15	0.1	-1.43	0.1539
listener	0.068	0.098	0.699	0.4849
speaker	0.16	0.079	2.08	0.03781
other	0.15	0.081	1.84	0.06583
attitude	0.075	0.12	0.605	0.5453
go	-0.078	0.046	-1.71	0.0884
drive	-0.16	0.047	-3.37	0.0008086
arrive	-0.2	0.048	-4.15	3.927e-05
n.believe	0.0084	0.058	0.147	0.8833
n.say	-0.011	0.058	-0.188	0.8507
n.quote	0.011	0.058	0.185	0.853
n.none	0.034	0.054	0.635	0.5259
c.believe:listener	-0.51	0.24	-2.18	0.03
c.believe:speaker	-0.16	0.2	-0.814	0.4158
c.believe:other	0.11	0.2	0.531	0.5957
c.say:listener	-0.38	0.28	-1.38	0.1675
c.say:speaker	-0.2	0.2	-0.985	0.3251
c.say:other	-0.11	0.21	-0.508	0.6118
c.quote:listener	0.15	0.23	0.625	0.5324
c.quote:speaker	0.055	0.22	0.257	0.7975
c.quote:other	0.027	0.22	0.122	0.9027
c.none:listener	0.064	0.24	0.268	0.7889
c.none:speaker	-5.3e-16	0.21	-2.56e-15	1
c.none:other	0.31	0.2	1.55	0.1229
c.believe:attitude	-0.23	0.19	-1.19	0.2348
c.say:attitude	-0.43	0.22	-1.99	0.0468
go:n.believe	0.018	0.12	0.148	0.8822
go:n.say	-0.033	0.12	-0.274	0.7845
go:n.quote	-0.00015	0.12	-0.00126	0.999
go:n.none	0.072	0.11	0.627	0.5306
drive:n.believe	-0.059	0.12	-0.49	0.6245
drive:n.say	-0.068	0.12	-0.545	0.5863
drive:n.quote	-0.076	0.12	-0.613	0.5398
drive:n.none	0.032	0.12	0.274	0.7842
arrive:n.believe	0.039	0.13	0.314	0.7533
arrive:n.say	-0.16	0.12	-1.27	0.2041
arrive:n.quote	-0.094	0.12	-0.757	0.4494
arrive:n.none	0.076	0.11	0.67	0.5032

Table 23: RoBERTa performance by syntactic environment and perspective-holder, perspectives nested in nested in *come* and compared to group mean

G.3 Combined models

Differences between RoBERTa and human scores were explored in a series of regression models.

Fixed effects (n=13632)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.76	0.0047	162	0
come	-0.23	0.012	-19.8	7.796e-86
c.roberta	0.015	0.02	0.787	0.431
n.roberta	-0.0033	0.02	-0.162	0.8716
go	-0.067	0.012	-5.59	2.357e-08
drive	-0.092	0.014	-6.61	3.984e-11
arrive	-0.11	0.015	-7.46	8.886e-14
n.roberta:go	0.017	0.037	0.472	0.6371
n.roberta:drive	-0.053	0.043	-1.21	0.2249
n.roberta:arrive	-0.094	0.047	-1.99	0.04693

Table 24: RoBERTa and human performance by verb, model nested in *come*

Fixed effects (n=13632)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.71	0.0085	83.5	0
come	-0.11	0.02	-5.42	6.109e-08
c.roberta	-0.017	0.019	-0.933	0.3507
c.annotated.corpus	-0.11	0.011	-9.92	4.268e-23
c.hard	-0.26	0.01	-24.9	1.708e-133
go	-0.033	0.021	-1.6	0.1107
drive	-0.046	0.023	-2	0.04577
arrive	-0.12	0.031	-3.74	0.0001831
n.roberta	-0.087	0.036	-2.43	0.01496
n.annotated.corpus	0.0028	0.015	0.184	0.8539
n.hard	-0.21	0.024	-8.75	2.469e-18
c.roberta:c.annotated.corpus	0.039	0.026	1.47	0.1411
go:n.roberta	0.055	0.066	0.829	0.4072
drive:n.roberta	-0.042	0.072	-0.583	0.5597
arrive:n.roberta	0.016	0.084	0.189	0.85
go:n.annotated.corpus	-0.05	0.021	-2.33	0.01999
go:n.hard	-0.029	0.04	-0.715	0.4744
drive:n.annotated.corpus	-0.064	0.025	-2.52	0.01178
drive:n.hard	-0.042	0.045	-0.937	0.3489
arrive:n.annotated.corpus	0.085	0.048	1.77	0.07666
arrive:n.hard	0.12	0.061	1.98	0.04805
n.roberta:n.annotated.corpus	-0.0061	0.037	-0.167	0.8675
n.roberta:n.hard	-0.22	0.07	-3.13	0.001759
go:n.roberta:n.annotated.corpus	-0.045	0.058	-0.771	0.4404
go:n.roberta:n.hard	0.027	0.13	0.207	0.8358
drive:n.roberta:n.annotated.corpus	-0.081	0.066	-1.22	0.2238
drive:n.roberta:n.hard	-0.095	0.14	-0.681	0.4961
arrive:n.roberta:n.annotated.corpus	-0.17	0.11	-1.56	0.1178
arrive:n.roberta:n.hard	-0.076	0.17	-0.462	0.644

Table 25: RoBERTa and human performance by verb and corpus, model nested in *come*

Fixed effects (n=13632)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.72	0.011	64.9	0
come	-0.13	0.027	-4.76	1.922e-06
c.roberta	-0.017	0.021	-0.8	0.4236
c.hard.spoken.corpus	-0.28	0.016	-17.5	1.089e-67
c.hard.written.corpus	-0.19	0.02	-9.36	9.063e-21
c.random.spoken.corpus	0.076	0.023	3.31	0.0009316
c.annotated.mod.corpus	-0.087	0.016	-5.55	2.925e-08
n.roberta	-0.16	0.048	-3.27	0.001076
n.hard.spoken.corpus	-0.042	0.05	-0.847	0.3972
n.hard.written.corpus	-0.28	0.034	-8.24	1.847e-16
n.random.spoken.corpus	0.069	0.035	1.98	0.04733
n.annotated.mod.corpus	0.028	0.023	1.26	0.2089
go	-0.066	0.027	-2.42	0.01542
drive	-0.075	0.031	-2.42	0.01559
arrive	-0.12	0.041	-2.97	0.003021
c.roberta:c.hard.spoken.corpus	-0.17	0.045	-3.76	0.0001689
c.roberta:c.hard.written.corpus	-0.18	0.061	-2.89	0.003908
c.roberta:c.random.spoken.corpus	-0.015	0.057	-0.26	0.7951
c.roberta:c.annotated.mod.corpus	-0.052	0.044	-1.19	0.2343
n.roberta:n.hard.spoken.corpus	-0.53	0.16	-3.35	0.0008145
n.roberta:n.hard.written.corpus	-0.2	0.1	-1.96	0.05016
n.roberta:n.random.spoken.corpus	-0.046	0.087	-0.53	0.5961
n.roberta:n.annotated.mod.corpus	-0.02	0.056	-0.365	0.7149
n.roberta:go	0.13	0.087	1.52	0.1297
n.roberta:drive	0.019	0.098	0.193	0.847
n.roberta:arrive	0.091	0.11	0.795	0.4267
n.hard.spoken.corpus:go	-0.38	0.091	-4.22	2.411e-05
n.hard.spoken.corpus:drive	-0.35	0.1	-3.41	0.0006463
n.hard.spoken.corpus:arrive	-0.049	0.11	-0.445	0.6567
n.hard.written.corpus:go	0.018	0.062	0.293	0.7696
n.hard.written.corpus:drive	-0.015	0.065	-0.224	0.8224
n.hard.written.corpus:arrive	0.13	0.082	1.55	0.1221
n.random.spoken.corpus:go	-0.1	0.046	-2.26	0.02414
n.random.spoken.corpus:drive	-0.081	0.059	-1.39	0.164
n.annotated.mod.corpus:go	-0.12	0.037	-3.25	0.001149
n.annotated.mod.corpus:drive	-0.13	0.041	-3.08	0.002064
n.annotated.mod.corpus:arrive	0.05	0.063	0.792	0.4286
n.roberta:n.hard.spoken.corpus:go	0.54	0.29	1.84	0.06537
n.roberta:n.hard.spoken.corpus:drive	0.27	0.33	0.824	0.4097
n.roberta:n.hard.spoken.corpus:arrive	0.31	0.35	0.887	0.3752
n.roberta:n.hard.written.corpus:go	-0.076	0.19	-0.394	0.6937
n.roberta:n.hard.written.corpus:drive	-0.23	0.2	-1.15	0.2497
n.roberta:n.hard.written.corpus:arrive	-0.17	0.23	-0.74	0.459
n.roberta:n.random.spoken.corpus:go	0.097	0.12	0.831	0.406
n.roberta:n.random.spoken.corpus:drive	-0.0026	0.14	-0.018	0.9857
n.roberta:n.annotated.mod.corpus:go	0.0045	0.098	0.0455	0.9637
n.roberta:n.annotated.mod.corpus:drive	-0.084	0.11	-0.786	0.4321
n.roberta:n.annotated.mod.corpus:arrive	-0.17	0.14	-1.18	0.24

Table 26: RoBERTa and human performance by modality, corpus, and verb, model nested in *come* and Written Random treated as baseline

Fixed effects (n=6657)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.79	0.0055	144	0
come	-0.2	0.013	-14.7	2.731e-48
c.roberta	0.0021	0.031	0.0682	0.9456
c.believe	0.02	0.023	0.87	0.3846
c.say	0.0074	0.024	0.31	0.7564
c.quote	0.024	0.024	0.999	0.3178
c.none	-0.079	0.023	-3.46	0.0005408
n.roberta	0.013	0.024	0.562	0.5742
go	-0.068	0.015	-4.66	3.232e-06
drive	-0.088	0.015	-5.9	3.846e-09
arrive	-0.09	0.015	-5.85	5.103e-09
n.believe	0.022	0.019	1.16	0.2455
n.say	0.0032	0.019	0.172	0.8637
n.quote	0.027	0.019	1.43	0.1539
n.none	0.015	0.017	0.84	0.4011
c.roberta:c.believe	0.019	0.076	0.254	0.7994
c.roberta:c.say	-0.025	0.079	-0.312	0.7552
c.roberta:c.quote	-0.14	0.081	-1.73	0.08389
c.roberta:c.none	-0.015	0.077	-0.195	0.8457
n.roberta:go	-0.02	0.049	-0.419	0.6751
n.roberta:drive	-0.08	0.05	-1.6	0.1087
n.roberta:arrive	-0.12	0.051	-2.34	0.01912
n.roberta:n.believe	-0.034	0.062	-0.557	0.5772
n.roberta:n.say	-0.035	0.062	-0.569	0.5691
n.roberta:n.quote	-0.037	0.062	-0.6	0.5486
n.roberta:n.none	-0.0017	0.058	-0.0294	0.9765
go:n.believe	0.012	0.039	0.313	0.754
go:n.say	-0.084	0.039	-2.16	0.03114
go:n.quote	0.0013	0.038	0.0329	0.9737
go:n.none	0.028	0.037	0.767	0.4431
drive:n.believe	-0.044	0.039	-1.15	0.2516
drive:n.say	0.0047	0.04	0.116	0.9078
drive:n.quote	-0.015	0.04	-0.373	0.7093
drive:n.none	0.022	0.038	0.577	0.5639
arrive:n.believe	0.043	0.04	1.06	0.2878
arrive:n.say	-0.058	0.04	-1.46	0.145
arrive:n.quote	-0.064	0.04	-1.6	0.1099
arrive:n.none	0.0011	0.037	0.0311	0.9752
n.roberta:go:n.believe	0.048	0.13	0.375	0.7079
n.roberta:go:n.say	0.094	0.13	0.725	0.4684
n.roberta:go:n.quote	0.041	0.13	0.324	0.7459
n.roberta:go:n.none	0.086	0.12	0.701	0.4832
n.roberta:drive:n.believe	0.028	0.13	0.217	0.8282
n.roberta:drive:n.say	-0.03	0.13	-0.224	0.823
n.roberta:drive:n.quote	-0.019	0.13	-0.141	0.8882
n.roberta:drive:n.none	0.053	0.13	0.423	0.6727
n.roberta:arrive:n.believe	0.039	0.13	0.291	0.7707
n.roberta:arrive:n.say	-0.056	0.13	-0.423	0.6725
n.roberta:arrive:n.quote	0.013	0.13	0.0942	0.9249
n.roberta:arrive:n.none	0.12	0.12	0.962	0.3358

Table 27: RoBERTa and human performance by syntactic environment and verb, model and evironment nested in *come*

Fixed effects (n=6657)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.79	0.0055	144	0
come	-0.13	0.0092	-14.4	1.87e-46
c.roberta	0.0021	0.032	0.0678	0.946
c.believe	0.02	0.023	0.864	0.3874
c.say	0.0074	0.024	0.308	0.7578
c.quote	0.024	0.024	0.993	0.3207
c.none	-0.079	0.023	-3.44	0.0005841
n.roberta	-0.00086	0.021	-0.04	0.9681
n.believe	0.019	0.017	1.14	0.2537
n.say	-0.009	0.017	-0.543	0.5875
n.quote	0.028	0.017	1.66	0.09772
n.none	0.023	0.016	1.44	0.151
c.roberta:c.believe	0.019	0.076	0.253	0.8006
c.roberta:c.say	-0.025	0.08	-0.31	0.7566
c.roberta:c.quote	-0.14	0.082	-1.72	0.08577
c.roberta:c.none	-0.015	0.077	-0.194	0.8466
n.roberta:n.believe	-0.017	0.056	-0.309	0.757
n.roberta:n.say	-0.026	0.055	-0.463	0.6437
n.roberta:n.quote	-0.024	0.055	-0.431	0.6662
n.roberta:n.none	0.023	0.053	0.428	0.669

Table 28: RoBERTa and human performance by syntactic environment, model and environment nested in *come*

Fixed effects (n=6657)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.79	0.0062	129	0
come	-0.19	0.017	-10.8	8.932e-27
c.roberta	0.033	0.057	0.586	0.5582
attitude.b	-0.022	0.029	-0.753	0.4514
listener.b	0.093	0.031	2.98	0.002853
anchor.b	-0.014	0.03	-0.469	0.6391
other.b	-0.069	0.025	-2.78	0.005398
n.roberta	0.013	0.023	0.559	0.5763
go	-0.067	0.015	-4.6	4.212e-06
drive	-0.089	0.015	-5.98	2.315e-09
arrive	-0.091	0.015	-6.03	1.696e-09
c.roberta:attitude.b	0.01	0.098	0.102	0.9185
c.roberta:listener.b	-0.15	0.1	-1.49	0.135
c.roberta:anchor.b	-0.15	0.1	-1.46	0.1432
c.roberta:other.b	0.053	0.083	0.643	0.5202
n.roberta:go	-0.016	0.048	-0.331	0.7408
n.roberta:drive	-0.075	0.049	-1.52	0.1289
n.roberta:arrive	-0.11	0.05	-2.21	0.02725

Table 29: RoBERTa and human performance by perspective-holder, perspective and model nested in *come*

Fixed effects (n=6657)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.79	0.0061	129	0
come	-0.13	0.015	-8.64	7.018e-18
c.roberta	0.039	0.057	0.686	0.4926
listener.b	0.045	0.034	1.32	0.1864
other.b	-0.066	0.026	-2.56	0.01052
anchor.b	-0.021	0.031	-0.678	0.4976
c.say	0.13	0.043	3.05	0.00231
c.believe	0.074	0.046	1.61	0.1073
c.quote	0.087	0.045	1.95	0.05123
c.none	-0.072	0.044	-1.63	0.1036
attitude.b	-0.026	0.038	-0.687	0.4924
n.roberta	-0.00086	0.021	-0.0403	0.9678
n.say	-0.009	0.016	-0.547	0.5846
n.believe	0.019	0.017	1.15	0.2501
n.quote	0.028	0.017	1.67	0.09519
n.none	0.023	0.016	1.45	0.1479
listener.b:c.say	-0.42	0.096	-4.36	1.292e-05
listener.b:c.believe	-0.24	0.083	-2.94	0.003271
listener.b:c.quote	-0.093	0.073	-1.28	0.2006
listener.b:c.none	-0.074	0.079	-0.947	0.3438
other.b:c.say	-0.25	0.065	-3.83	0.0001273
other.b:c.believe	-0.13	0.067	-1.97	0.04865
other.b:c.quote	-0.29	0.065	-4.4	1.075e-05
other.b:c.none	-0.072	0.058	-1.24	0.2156
anchor.b:c.say	0.18	0.081	2.19	0.02845
anchor.b:c.believe	0.14	0.078	1.8	0.07142
anchor.b:c.quote	0.14	0.085	1.65	0.09932
anchor.b:c.none	0.24	0.081	2.94	0.003329
c.say:attitude.b	-0.22	0.066	-3.35	0.0008274
c.believe:attitude.b	-0.017	0.064	-0.265	0.7913
c.roberta:listener.b	-0.16	0.11	-1.46	0.145
c.roberta:other.b	0.049	0.086	0.568	0.5703
c.roberta:anchor.b	-0.18	0.1	-1.76	0.07794
c.roberta:c.say	-0.14	0.14	-0.998	0.3182
c.roberta:c.believe	-0.074	0.15	-0.485	0.628
c.roberta:c.quote	-0.22	0.15	-1.46	0.1452
c.roberta:c.none	-0.17	0.15	-1.15	0.2521
c.roberta:attitude.b	-0.055	0.13	-0.434	0.6645
n.roberta:n.say	-0.026	0.055	-0.466	0.6412
n.roberta:n.believe	-0.017	0.055	-0.312	0.7552
n.roberta:n.quote	-0.024	0.055	-0.435	0.6639
n.roberta:n.none	0.023	0.052	0.431	0.6666
c.roberta:listener.b:c.say	0.19	0.32	0.6	0.5489
c.roberta:listener.b:c.believe	-0.2	0.28	-0.706	0.48
c.roberta:listener.b:c.quote	0.21	0.24	0.849	0.396
c.roberta:listener.b:c.none	0.15	0.26	0.589	0.5556
c.roberta:other.b:c.say	0.37	0.22	1.69	0.0905
c.roberta:other.b:c.believe	0.47	0.22	2.08	0.03774
c.roberta:other.b:c.quote	0.25	0.22	1.15	0.2486
c.roberta:other.b:c.none	0.46	0.2	2.34	0.01942

Table 30: RoBERTa and human performance by perspective-holder and syntactic environment, perspective and model nested in *come* and Speaker treated as baseline

Fixed effects (n=6657)	Estimate	Std. Error	z value	Pr(>|z|)
c.roberta:other.b:c.say	0.37	0.22	1.69	0.0905
c.roberta:other.b:c.believe	0.47	0.22	2.08	0.03774
c.roberta:other.b:c.quote	0.25	0.22	1.15	0.2486
c.roberta:other.b:c.none	0.46	0.2	2.34	0.01942
c.roberta:anchor.b:c.say	0.074	0.27	0.272	0.7856
c.roberta:anchor.b:c.believe	0.06	0.26	0.23	0.8183
c.roberta:anchor.b:c.quote	-0.21	0.29	-0.736	0.4616
c.roberta:anchor.b:c.none	-0.24	0.27	-0.873	0.3827
c.roberta:c.say:attitude.b	-0.063	0.22	-0.286	0.7745
c.roberta:c.believe:attitude.b	-0.023	0.21	-0.109	0.9133

Table 31: RoBERTa and human performance by perspective-holder and syntactic environment, perspective and model nested in *come* and Speaker treated as baseline

Fixed effects (n=6657)	Estimate	Std. Error	z value	Pr(>|z|)
c.roberta:listener:c.say	0.095	0.29	0.322	0.7475
c.roberta:listener:c.believe	-0.2	0.25	-0.819	0.4128
c.roberta:listener:c.quote	0.33	0.25	1.35	0.178
c.roberta:listener:c.none	0.31	0.25	1.24	0.2143
c.roberta:other:c.say	0.23	0.23	1.04	0.2993
c.roberta:other:c.believe	0.32	0.21	1.53	0.1272
c.roberta:other:c.quote	0.37	0.23	1.59	0.1111
c.roberta:other:c.none	0.56	0.21	2.64	0.008366
c.roberta:speaker:c.say	-0.059	0.22	-0.272	0.7856
c.roberta:speaker:c.believe	-0.048	0.21	-0.23	0.8183
c.roberta:speaker:c.quote	0.17	0.23	0.736	0.4616
c.roberta:speaker:c.none	0.19	0.22	0.873	0.3827
c.roberta:c.say:attitude	-0.11	0.23	-0.479	0.6318
c.roberta:c.believe:attitude	-0.066	0.2	-0.323	0.7464

Table 32: RoBERTa and human performance by perspective-holder and syntactic environment, perspective and model nested in *come* and perspectives compared to group mean

Fixed effects (n=6657)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.79	0.0057	140	0
come	-0.12	0.011	-11	6.28e-28
c.roberta	-0.054	0.041	-1.31	0.1916
listener	0.052	0.031	1.69	0.09179
other	-0.036	0.026	-1.38	0.1667
speaker	0.017	0.025	0.678	0.4976
c.say	-0.013	0.028	-0.458	0.6467
c.believe	0.024	0.025	0.966	0.334
c.quote	0.068	0.034	2.01	0.04476
c.none	-0.0058	0.032	-0.18	0.857
attitude	0.072	0.039	1.82	0.06935
n.roberta	-0.00086	0.021	-0.0403	0.9678
n.say	-0.009	0.016	-0.547	0.5846
n.believe	0.019	0.017	1.15	0.2501
n.quote	0.028	0.017	1.67	0.09519
n.none	0.023	0.016	1.45	0.1479
listener:c.say	-0.48	0.088	-5.42	6.315e-08
listener:c.believe	-0.31	0.075	-4.13	3.731e-05
listener:c.quote	-0.19	0.074	-2.53	0.01154
listener:c.none	-0.25	0.075	-3.32	0.0009119
other:c.say	-0.34	0.068	-5.05	4.523e-07
other:c.believe	-0.22	0.063	-3.44	0.000587
other:c.quote	-0.34	0.069	-4.93	8.419e-07
other:c.none	-0.25	0.062	-3.97	7.121e-05
speaker:c.say	-0.14	0.065	-2.19	0.02845
speaker:c.believe	-0.11	0.062	-1.8	0.07142
speaker:c.quote	-0.11	0.068	-1.65	0.09932
speaker:c.none	-0.19	0.065	-2.94	0.003329
c.say:attitude	-0.32	0.069	-4.67	3.068e-06
c.believe:attitude	-0.13	0.061	-2.05	0.03999
c.roberta:listener	0.016	0.1	0.154	0.8778
c.roberta:other	0.19	0.086	2.15	0.03129
c.roberta:speaker	0.15	0.083	1.76	0.07794
c.roberta:c.say	-0.028	0.092	-0.306	0.7596
c.roberta:c.believe	-0.013	0.082	-0.159	0.8735
c.roberta:c.quote	-0.21	0.11	-1.87	0.06177
c.roberta:c.none	-0.14	0.11	-1.3	0.1952
c.roberta:attitude	0.013	0.13	0.0978	0.9221
n.roberta:n.say	-0.026	0.055	-0.466	0.6412
n.roberta:n.believe	-0.017	0.055	-0.312	0.7552
n.roberta:n.quote	-0.024	0.055	-0.435	0.6639
n.roberta:n.none	0.023	0.052	0.431	0.6666

Table 33: RoBERTa and human performance by perspective-holder and syntactic environment, perspective and model nested in *come* and perspectives compared to group mean

H ProSPer examples with lowest human mean accuracy

H.1 Annotated subset

1. **Human mean: 0%**
   RoBERTa score: 1.6%
   Target: *come*
   "You should go to Jinyang too," she answered coldly. "There are already too many people," he answered. "My sister's house is full. And there's no work in Jinyang." "What do you want?" "I'll leave Lan Lan in Jinyang– none of the schools are open now, anyway– and my parents can take care of her." "And you?" "I thought I might **come** to Beijing." He said this weakly, and the next sentence was even more feeble. "There's nothing left here." Chrysanthemum was silent for a long time. Then she said, "Let me think about it." She hung up and suddenly realized no other plan of action was possible.
2. **Human mean: 0%**
   RoBERTa score: 39%
   Target: come
   As Luc lived alone in Lyon, his son tried hard to **come** there every week.
3. **Human mean: 0%**
   RoBERTa score: 1.0%
   Target: come
   Charles, I would like to go on record, the tux, not my idea. No, no. But they said if you want to **come** behind the scenes on the number one show in America, you got ta succumb to the dress code.
4. **Human mean: 0%**
   RoBERTa score: 51%
   Target: walk
   CORSON: Did he say where he was going when he got out? PAUL: Yes. CORSON: Where? PAUL: Paris. OSCAR: see what I mean? Stupid. PAUL: He just wanted to think about it. The Eiffel Tower, he wanted to see, aperitifs, he wanted to drink, the Follies Brassiere, he wanted to go to. Beautiful buildings, cafs, music, and women. I told him what I knew. Then he showed me a map. CORSON: A map? PAUL: He said I could **walk** to Paris. CORSON: From where? PAUL: From here. OSCAR: Aw, man. PAUL: He had it all worked out. You go up through Laos, then into Burma, some other country I forget, then India, Iran, Turkey, and Greece. The rest, he said, is easy.
5. **Human mean: 0%**
   RoBERTa score: 82%
   Target: walking
   It's hard to keep calm and carry on, not let anyone see me at my most vulnerable, when I'm away from home and I don't have hugs in the kitchen with my mum or chats by the fire with my dad to see me through the day. Sometimes it's just a case of putting on the saddest Taylor Swift song while you're **walking** home from the laundrette, crying quietly.
6. **Human mean: 0%**
   RoBERTa score: 3.6%
   Target: arrive
   Kristen wrote to him in a joyful, loopy scrawl from Middlebury on a postcard which bore a bright blue picture of the school's ski team. The card said Sophie wanted Frank to **arrive** home early because she was giving him a surprise graduation party. Kristen was meddling in her parents' lives with characteristic thoughtfulness, not wanting Frank to disappoint Sophie by being late, not wanting him to be so surprised himself that he couldn't enjoy the party.
7. **Human mean: 0%**
   RoBERTa score: 2.1%
   Target: drive
   So where does your husband live? I asked Marietta. She sat back in the armchair. On top of the hill, before you turn to 248, there's a white trailer, with many pots of flowers in the windows. You are welcome to stop by. To see him? Well, no, there's another trailer, a pink one, and that's mine, way back in the yard. You think I could just **drive** in and talk to you all? Your trailer doesn't have the flowers? No, I have cats. They're always sunbathing in the windows. Why wouldn't you stop by? If you usually don't even talk to each other? Oh, we do, I just hope we won't. We have a kid to take care of, so that keeps us pretty close.
8. **Human mean: 0%**
   RoBERTa score: 0.9%
   Target: arrive
   There was also a charming lady in her late thirties or perhaps early forties who was secretary to a member of Parliament, She was responsible for my enjoying a weekend at the country home of well-to-do friends of hers. She remarked at dinner once that she had an invitation to spend the weekend with friends but was unable to accept. So, playing the role of the brash American, I suggested that she tell them I would **arrive** as substitute. She did, and I did. Her friends entertained me extravagantly, I thought, and took me around to meet their friends, who also felt obliged to entertain me. One of them later visited us in Paris, and in return for his and his wife's hospitality I took him to dinner in Paris and then to the opera.
9. **Human mean: 0%**
   RoBERTa score: 22%
   Target: come
   We were still cheerful when we got back to Johannesburg next evening, in spite of the weather. It was one of those bitter Transvaal nights, with a sharp, icy drizzle. We arrived at the house about eight. For a moment we both thought we had **come** up the wrong street, because there on the lawn was a red tent – one of those long low ones they call patrol tents – tied against the kaya.
10. **Human mean: 0%**
    RoBERTa score: 10%

Target: came
YDSTIE: When you went to Rome, you brought along the beginnings of a World War II novel you were working on, but you got distracted. Tell us what happened? Mr-DOERR: Yes, or overwhelmed perhaps is the better word. I **came** there with about 50 pages in notes on this novel thinking that that would be my next project. I had just published my second book, a novel, and was trying to get a new project off the ground. And usually, I think I learned there that I prefer a very insulated environment where everything is pretty familiar for me to be able to enter and imagine this space and start writing fiction. And Rome invaded me on all sides, even from the first day.

H.2 Random subset

1. **Human mean: 0%**
   RoBERTa score: 10%
   Target: came
   and we'd go up to the Adirondacks and camp and it was so you know pick your own uh blueberries and make blueberry pancakes for breakfast uh also go ahead oh yes yes yes i **came** i
2. **Human mean: 0%**
   RoBERTa score: 87%
   Target: drove
   Dorians and Romans After the disaster, Mycenean Greeks from the Peloponnese moved in to control what remained of the Minoan settlements — they may even have precipitated the destruction. Around 1200 b.c., Dorian invaders from the Balkans **drove** south through the Greek mainland, the Aegean islands, and across to Crete. Many coastal dwellers migrated to remote mountain villages in order to escape their enemies. Others embarked on an overseas exodus that took them around the Mediterranean Sea. The island did not get directly involved in Greece's Persian and Peloponnesian Wars, however it became well-known as a valuable source of brave and energetic mercenaries.
3. **Human mean: 0%**
   RoBERTa score: 22%
   Target: coming
   Robe's collection of New Mexico legends contains thirty-four variants of the devil-at-the-dance tale. Although the legends in this collection are from rural northern New Mexico, collected in the 1950s and 1960s, we find contemporary versions of the devil-at-the-dance tales in south Texas and in Baja California from the 1980s. Limón and Herrera-Sobek discuss versions of the tale circulating in nightclubs and discotheques among urbanized young people. Of course, not everyone believes such stories. Martin's book contains a story by a man born in 1904, who says his friend played a trick on his community in Tucson by **coming** to a dance dressed in black, with a fake rooster foot. Eventually someone noticed his foot and yelled, "The Devil! The Devil!" The narrator says he was there when his friend played the trick, so he doesn't believe in the legend (50). References De Leon 1982; Glazer 1984,1994; Herrera-Sobek 1988; Limón 1994; Martin 1983; Robe 1951; Robe ed. 1980; West 1988
4. **Human mean: 0%**
   RoBERTa score: 77%
   Target: come
   supposed to convince the market that all was well, so eventually interest rates could **come** down. The trouble was that all that austerity was a hard sell politically–especially because the economy was going into a nasty recession
5. **Human mean: 0%**
   RoBERTa score: 52%
   Target: going
   uh above or around or uh from other places in in well if he got out of line too much from the old school they'd just knock him off i mean they're not **going** they're not
6. **Human mean: 0%**
   RoBERTa score: 71%
   Target: go
   um-hum hopefully next year get to **go** back because uh a lot of the family has not uh on her side my mother's side has not seen you know my daughter so
7. **Human mean: 0%**
   RoBERTa score: 56%
   Target: went
   Well, they **went** through, they broke a soldier's leg, you could hear his moaning. Then they
8. **Human mean: 17%**
   RoBERTa score: 0.8%
   Target: went
   contained both cocaine and an extract from the cola nut. The cocaine **went** out when it was declared a controlled substance early in this century. Interestingly,
9. **Human mean: 17%**
   RoBERTa score: 82%
   Target: come
   intent of attracting people to treatment because patients do not **come** to the emergency department with the intention of receiving substance abuse treatment.

I ProSPer examples with lowest RoBERTa accuracy

I.1 Annotated subset

1. **Human mean: 10%**
   RoBERTa score: 0.2%
   Target: drive
   "You want to turn back after lunch?" she offered. Hisako looked up from her snow picnic hesitantly. "I know how badly you want to go to the hot springs..." "Not so bad I'll drag you there when it's making you feel awful." "Really?" "It's a holiday, right?" "You won't be mad at me later?" "What do you take me for?" Megumi asked. "We can **drive** down to the public one. It's not as nice, but it'll be warm." "We've been walking for over three hours," Hisako looked at her watch. "It's getting kind of darker, don't you think." "It'll be faster going downhill," Megumi promised. sniffing and Megumi didn't know if it was only her nose, or if she was crying.
2. **Human mean: 50%**
   RoBERTa score: 0.6%
   Target: arrive
   ASSURAS: When you immigrated to the United States, you were 19 years old and you had huge expectations, didn't you? Mr-McCOURT: Well, this w– this was the golden land for– for most people who come here, so I– I– I thought I'd just– I– I thought I'd **arrive** here, get off the boat and– and go marching up Fifth Avenue like Fred Astaire and Ginger Rogers, and that music would play in the background because all the movies had music in the background. I thought that– that's how it was in America. ASSURAS: It didn't happen that way, though, did it? Mr-McCOURT: No, it didn't because somebody has to take out the garbage. And that's– that's the kind of job I got right away.
3. **Human mean: 0%**
   RoBERTa score: 0.9%
   Target: arrive
   "There was also a charming lady in her late thirties or perhaps early forties who was secretary to a member of Parliament, She was responsible for my enjoying a weekend at the country home of well-to-do friends of hers. She remarked at dinner once that she had an invitation to spend the weekend with friends but was unable to accept. So, playing the role of the brash American, I suggested that she tell them I would **arrive** as substitute. She did, and I did. Her friends entertained me extravagantly, I thought, and took me around to meet their friends, who also felt obliged to entertain me. One of them later visited us in Paris, and in return for his and his wife's hospitality I took him to dinner in Paris and then to the opera."
4. **Human mean: 0%**
   RoBERTa score: 1.0%
   Target: come
   Charles, I would like to go on record, the tux, not my idea. No, no. But they said if you want to **come** behind the scenes on the number one show in America, you got ta succumb to the dress code.
5. **Human mean: 27%**
   RoBERTa score: 1.6%
   Target: come
   The clout list shows an Anthony Scarpelli sought a city job in 1994 with help from Terry Teele, a former Daley aide. Scarpelli said he's not sure if that's him or his father, also a city worker on disability leave. Scarpelli, a cousin of the late mobster Gerald Scarpelli, said he also helped run a patronage army for Carmen Iacullo and Anthony Pucillo, two former transportation department officials. "I want to **come** back to work," said Scarpelli , who settled four previous cases for $28,000.
6. **Human mean: 0%**
   RoBERTa score: 1.9%
   Target: come
   "You should go to Jinyang too," she answered coldly. "There are already too many people," he answered. "My sister's house is full. And there's no work in Jinyang." "What do you want?" "I'll leave Lan Lan in Jinyang– none of the schools are open now, anyway– and my parents can take care of her." "And you?" "I thought I might **come** to Beijing." He said this weakly, and the next sentence was even more feeble. "There's nothing left here." Chrysanthemum was silent for a long time. Then she said, "Let me think about it." She hung up and suddenly realized no other plan of action was possible.
7. **Human mean: 0%**
   RoBERTa score: 2.1%
   Target: drive
   So where does your husband live? I asked Marietta. She sat back in the armchair. On top of the hill, before you turn to 248, there's a white trailer, with many pots of flowers in the windows. You are welcome to stop by. To see him? Well, no, there's another trailer, a pink one, and that's mine, way back in the yard. You think I could just **drive** in and talk to you all? Your trailer doesn't have the flowers? No, I have cats. They're always sunbathing in the windows. Why wouldn't you stop by? If you usually don't even talk to each other? Oh, we do, I just hope we won't. We have a kid to take care of, so that keeps us pretty close.
8. **Human mean: 70%**
   RoBERTa score: 2.4%
   Target: come
   "They talked about Harry Logan and Chatham and a new outbreak of Lyme disease that everyone was worrying about. Elliott said they knew a young girl about Grace's age who'd caught it and her life had been completely wrecked. Connie darted a look at him and he flushed a little and quickly changed the subject. As soon as the meal was over, Grace said she was tired and would they mind if she went to bed. Annie said she would **come** too but Grace wouldn't let her."
9. **Human mean: 10%**
   RoBERTa score: 2.4%

Target: came
The house of my paternal grandparents in central Switzerland: that was the "other" home of my childhood. And at the same time, it was another world. Back then, it took a long time to get there, and that universe had nothing to do with my life in Paris: it was completely surrounded by countryside, and I was rid of the school I hated. From early on, I **came** there alone. I was reunited with my Roman cousins and experienced a completely different lifestyle. I smelled different odors. I saw new lights both outside and inside the home where, once evening fell, the large lamps spread out on various pieces of furniture or hanging above the big dinner table made centers of luminous warmth.

10. **Human mean: 50%**
RoBERTa score: 2.6%
Target: arrive
I had a snapped torsion spring and needed it fixed. I also had an ancient Genie screw drive motor that I'm pretty sure the neighbors 3 blocks away could hear each time we opened our garage door. I called Art who answered right away, we talked about the cost to replace the springs and some options for openers. He said they'd **arrive** the next day between 9-10 after they finished a job they had that morning.

I.2 Random subset

1. **Human mean: 20%**
RoBERTa score: 0.6%
Target: arrive
More than 30,000 Cubans tried to cross shark-infested waters to Florida on improvised rafts. Facing a dramatic influx of Cubans, President Clinton abolished the US policy of automatic asylum to Cuban refugees, placing them in a makeshift tent settlement in Guantánamo Bay Naval Base. In the opinion of many, Cuba is an isolated socialist dinosaur. Yet it— and its aging leader— soldier on against all odds. The US embargo, denounced by an ever-increasing majority in the United Nations, is spurred on by vehement anti-Castro lobbying by Cubans in southern Florida. Still, changes in the status quo seem likely. But nearly four decades after the Cuban missile crisis, the world continues to wait, wondering whether a lifting of US trade and tourism sanctions or a change in leadership in Cuba will **arrive** first.
2. **Human mean: 17%**
RoBERTa score: 0.8%
Target: went
contained both cocaine and an extract from the cola nut. The cocaine **went** out when it was declared a controlled substance early in this century. Interestingly,
3. **Human mean: 40%**
RoBERTa score: 1.3%
Target: goes
that by the conservatives' lights, the bill spends too much and cuts taxes too little. The NYT even **goes** out of its piece saying that the vote revealed a severe split in the Republican Party that may be hard to patch up.
4. **Human mean: 20%**
RoBERTa score: 1.4%, target: went
Commercial enterprise, military power, and religious fervor **went** together. More than the divinity of Shiva, the 11th-century Temple of Brihadisvara boasts architecture celebrating the victory of
5. **Human mean: -**[16]
RoBERTa score: 2.1%
Target: drove
yeah well in um in Dallas um i don't know if you heard about the killing where the guy **drove** into Luby's and the story was uh
6. **Human mean: 20%**
RoBERTa score: 2.7%
Target: walks
. More generally, each organism has traits that are aected by many genes, the polygeny discussed above, and each gene aects many traits, the pleiotropy alluded to above. It is interesting to note that were organisms to evolve to a position below but near the biological reality that is the proper analogue of the Ksat phase transition, such a location might well achieve the gradualism and capacity to persistently evolve that Darwin noted and that we observe. Both the gradualism and capacity to evolve are related to the number of alternative assignments of true or false to the V variables that satisfy the Ksat normal disjunctive form. If there are connected pathways from one such assignment via -Hamming-mutant neighboring assignments that all satisfy the normal disjunctive form, then adaptive **walks** via alternatives genotypes are available, all of which roughly generate the same organism. Gradualism is achieved. Polygeny and pleiotropy tune landscape ruggedness and deformability, which tune coevolutionary dynamics, perhaps to a self-organized critical state of an ecosystem. There are, in short, dimly understood laws that allow the coevolutionary construction of accumulating complexity. And it appears that such coevolution typically is self-organized critical. The NK coevolutionary model is not the only example of a model exhibiting self-organized critical behavior of model ecosystems. Bak and colleagues, Ricard Solé, and others have created elegant models aiming in the same direction. In particular, Solé's model comes closest to fitting the actual slopes of the observed power laws, whi
7. **Human mean: 80%**
RoBERTa score: 3.8%
Target: drive
yes they are well most of them you're right but then i **drive** through Riodosa and i think well not all of them have it
8. **Human mean: 100%**
RoBERTa score: 3.9%

[16] Removed due to violent content.

Target: go
we've **go** t all of our trees from Calloway's and luckily they'll take them back any time for any reason if they die yeah i usually

9. **Human mean: 40%**
   RoBERTa score: 6.0%
   Target: went
   out the discerning sense of discriminating. Disinterested **went** next, the eulogy continued, becoming a posh synonym for uninterested , because
10. **Human mean: 20%**
    RoBERTa score: 6.3%
    Target: coming
    to slam a competitor via the most unprofessional means. The kind of SLATE I was to **coming** enjoy reading was one that would rip the Wall Street Journal apart using facts, ideas, and finely honed reasoning. There is no place in that