Investigating representations of verb bias in neural language models

Robert D. Hawkins[*1], **Takateru Yamakoshi**[*1,2], **Thomas L. Griffiths**[1], **Adele E. Goldberg**[1]
[1]Princeton University, [2]University of Tokyo
{rdhawkins,takateru,tomg,adele}@princeton.edu

Abstract

Languages typically provide more than one grammatical construction to express certain types of messages. A speaker's choice of construction is known to depend on multiple factors, including the choice of main verb – a phenomenon known as *verb bias*. Here we introduce DAIS, a large benchmark dataset containing 50K human judgments for 5K distinct sentence pairs in the English dative alternation. This dataset includes 200 unique verbs and systematically varies the definiteness and length of arguments. We use this dataset, as well as an existing corpus of naturally occurring data, to evaluate how well recent neural language models capture human preferences. Results show that larger models perform better than smaller models, and transformer architectures (e.g. GPT-2) tend to out-perform recurrent architectures (e.g. LSTMs) even under comparable parameter and training settings. Additional analyses of internal feature representations suggest that transformers may better integrate specific lexical information with grammatical constructions.

1 Introduction

When we use language, we are often faced with a choice between several possible ways of expressing the same message. For example, in English, to express an event of intended or actual transfer between two animate entities, one option is the *double-object* (DO) construction, in which two noun phrases follow the verb. Alternatively, the same content can be expressed using the *prepositional dative* (PO) construction.

(1) a. Ava gave him something. *DO*
b. Ava gave something to him. *PO*

Speakers' preferences for one or the other construction depend on multiple factors, including the length and definiteness of the arguments (Oehrle, 1976; Arnold et al., 2000; Wasow, 2002; Bresnan, 2007). One particularly subtle factor is the lexical *verb bias*. While some verbs readily occur in either construction, others have strong preferences for one over the other (Levin, 1993):

(2) a. ?Ava said him something. *DO*
b. Ava said something to him. *PO*

Decades of work in linguistics and psychology has investigated how humans learn these distinctions (Gropen et al., 1989; Perfors et al., 2010; Barak et al., 2014; Goldberg, 2019). Yet, as deep neural networks have achieved state-of-the-art performance across many tasks in natural language processing, little is known about the extent to which they have acquired similarly fine-grained preferences. Although neural language models robustly capture certain types of grammatical constraints, e.g., subject-verb agreement and long distance dependencies (Linzen and Baroni, 2021; Manning et al., 2020), they continue to struggle with other aspects of syntax, including argument structure (e.g. Warstadt et al., 2019). Verb biases provide a particularly interesting testbed. Successfully predicting these psycholinguistic phenomena requires the integration of specific lexical information with representations of higher-level grammatical structures, with implications for understanding differential performance between models on other tasks.

In the current work, we take an analytic and comparative approach. First, we introduce the *DAIS* (*D*ative *A*lternation and *I*nformation *S*tructure) dataset, containing 50K human preference judgments for 5K sentence pairs, using 200 unique verbs. These empirical judgments indicate that verb bias preferences are highly gradient in practice (Ryskin et al., 2017; Ambridge et al., 2018), rather than belonging to binary "alternating" and "non-alternating" classes, as commonly assumed. Second, we evaluate the predictions of a variety of neural models, including both recurrent architectures and transformers, and analyze their internal

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4653–4663,
November 16–20, 2020.

states to understand what drives differences in performance. Finally, we evaluate our models on natural production data from the Switchboard corpus, finding that transformers achieve similar classification accuracy as prior work using hand-annotated features ($\sim 93\%$; Bresnan et al., 2007).

2 Related Work

Several recent studies have investigated how neural language models represent the dative alternation. Kann et al. (2018) constructed a corpus of verbs in common alternations, including the dative, and showed that a degree of information about acceptability is decodable directly from embeddings of the verb. However, acceptability was not based on empirical data and verb bias was treated as a binary variable, preventing an analysis of gradient effects. Kelly et al. (2020) found that DO constructions are separable from non-DO constructions in high-dimensional sentence embeddings (including BERT), but did not investigate verb bias. Futrell and Levy (2019) confirmed that recurrent neural networks (RNNs) show human-like sensitivity to several other important aspects of gradience in dative alternations, including the length and definiteness of arguments. However, they included only 16 verbs, all considered "alternating." Additionally, in these studies, a limited range of neural models were considered, leaving it unclear exactly how predictions may depend on architectural choices, model size, and training regime.

3 The DAIS dataset

The DAIS dataset contains 50,136 human preference judgments for 5,000 sentence pairs, constructed as follows. First, to obtain a large and heterogeneous set of verbs, we collected the 100 most frequent verbs influentially classified by Levin (1993) as alternating (i.e. acceptably appearing in both PO and DO constructions), as well as the 100 most frequent verbs classified as "non-alternating" (appearing only in the PO construction). This set contains most of the verbs examined in prior corpus-based analyses (see Sec. 5). For each verb, we generated DO and PO sentences across 5 different conditions, manipulating the length and definiteness of the recipient argument (see ex. 3).

(3) a. Ava gave *him* a book.
b. Ava gave *the man* a book.
c. Ava gave *a man* a book.
d. Ava gave *the man from work* a book.
e. Ava gave *a man from work* a book.

Finally, to obtain a range of distinct items in each condition, we created 5 plausible theme arguments for each verb, including 2 definite and 3 indefinite, for a total of 5,000 alternation pairs.

We collected judgments from 1011 participants on Amazon Mechanical Turk. Each participant was shown 50 dative alternation pairs (DO vs. PO) using unique verbs, balanced across the possible recipient and theme conditions. On each trial, participants used a continuous slider to indicate the strength of their preference for the DO or the PO, with the midpoint used to indicate they were "about the same" (see Appendix A for details)[1].

4 Results

4.1 Characterizing human judgments

We begin by characterizing benchmark patterns of human judgments in our dataset. First, we examine the degree of gradience in DO preference across verbs. Traditionally, verbs have been grouped into binary "classes": alternating verbs which appear freely in both constructions, and non-alternating verbs which are only acceptable in one (Levin, 1993). While verbs in the "alternating" class were indeed rated more acceptable on average in the DO than "non-alternating" verbs ($b = -15.0, t = -46.5, p < 0.001$), there was

[1]Our procedure and behavioral analysis plan were preregistered at **https://osf.io/rtzv4** and we have released all data and analysis code at **https://github.com/taka-yamakoshi/neural_constructions**.

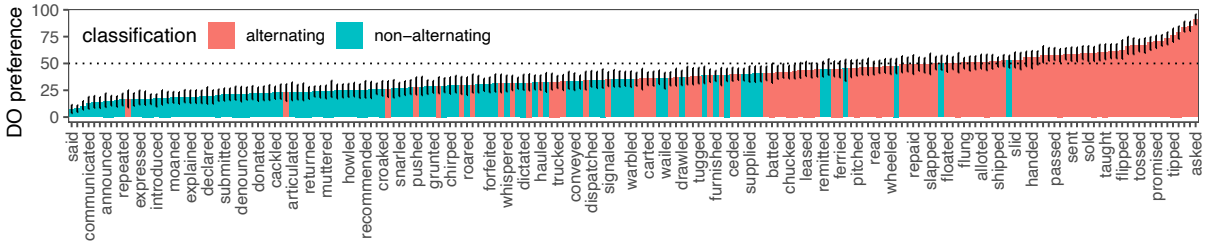

Figure 1: Human judgments across 200 verbs, pronoun recipients only. Classification from Levin (1993).

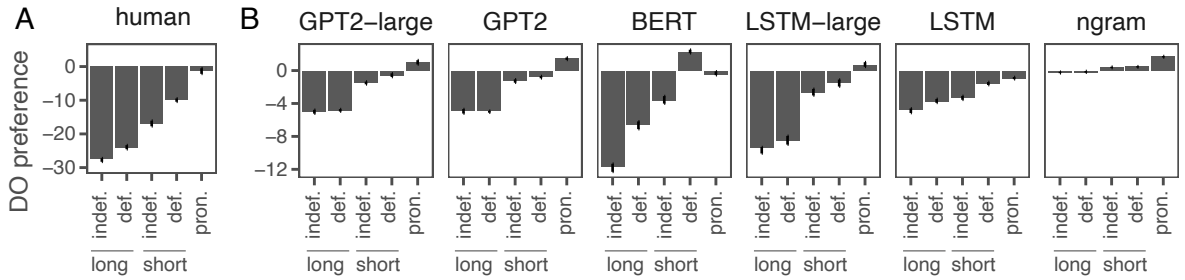

Figure 2: Average (A) human and (B) model DO preferences for "alternating" verbs, as recipient argument varies in length and definiteness. For "non-alternating" verbs and effects of theme definiteness and, see Fig. S1 in Appendix.

substantial overlap between the two classes, confirming the need to verify introspective classifications with human judgments. Moreover, we found that verbs fell along a continuous spectrum of acceptability in the DO (Lau et al., 2017; Gibson and Fedorenko, 2013, see Fig. 1; individual responses are shown in Supplemental Fig. S2). To measure the stability of this ranking across participants, we repeatedly split the dataset in half, measured the mean judgment for each verb across recipient conditions, and took the Spearman correlation between the two halves. Across 100 splits, we found an average correlation of $r = 0.95$, which also serves as a noise ceiling for our model comparison.[2]

Second, we examine human sensitivity to the length and definiteness of the arguments (Fig. 2A). Consistent with previous findings (Wasow, 2002; Futrell and Levy, 2019), participants more strongly preferred the double-object construction when the recipient was shorter ($b = 16.3, t = 21.1, p < 0.001$) and definite ($b = 3.9, t = 14.4, p < 0.001$), and when the theme was indefinite ($b = 2.2, t = 11.0, p < 0.001$; see Appendix B for more details). These effects were roughly additive: although longer recipient arguments rarely occur in the DO construction, we nonetheless found a preference for long definite arguments compared to long indefinites. Similar effects were found when limiting analysis to only "non-alternating" verbs (see Fig. S1 in Appendix). Indeed, "non-alternating" verbs with short, pronoun recipients were judged to be more acceptable in the double-object construction than "alternating" verbs with long, indefinite recipients, highlighting the interplay between verb biases and information structure.

[2]One concern is that gradience is an artifact of using a continuous slider (Armstrong et al., 1983; Yang, 2008). Recent work (e.g. Lau et al., 2017) has addressed this concern by examining the histograms of ratings on different measures, finding higher similarity to gradient control tasks than binary control tasks. Still, it is important to note that gradient judgments are compatible with categorical grammars due to multiple binary factors or individual differences (Schütze, 2011).

	# Layers	Hidden dim.	# params	Data (# tokens)
Ngram	-	-	-	English Wikipedia subset (80M)
LSTM	2	650	0.17M	English Wikipedia subset (90M)
LSTM-large	2	1024	1.04B	One Billion Word Benchmark(800M)
BERT	12	768	110M	BooksCorpus (800M) and English Wikipedia(2.5B)
GPT2	12	768	117M	WebText(8B in estimate)
GPT2-large	36	1280	774M	WebText(8B in estimate)

Table 1: Details of each model we consider.

4.2 Comparing model predictions

Next, we evaluated the performance of several pre-trained neural language models (see Table 1) against the fine-grained human judgments in DAIS. We included two recurrent architectures of different sizes: the 2-layer LSTM model from Gulordava et al. (2018), which has been used for a variety of previous syntactic evaluations, as well as the larger 1B-parameter "BIG LSTM+CNN" (Jozefowicz et al., 2016). We also included several transformer architectures, including BERT (Devlin et al., 2018), and two sizes of GPT-2 (Radford et al., 2019). These choices allow us to explore both effects of architecture as well as size and training regime. As a baseline, we included a 5-gram model and interpolated using the methods described in Heafield et al. (2013).

For the LSTM and GPT2 architectures, we calculated sentence probabilities by taking the sum of the surprisal of each word, conditioning on all of the preceding words. For BERT, which uses bidirectional context, we used the surprisal of each word conditioned on the full context (Wang and Cho, 2019). We then measured the models' relative preference for the DO construction by taking the likelihood ratio of the two sentences.

We began by examining how each model cap-

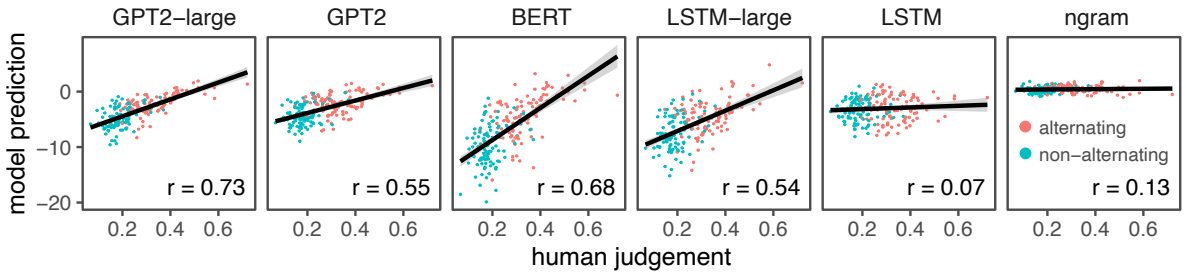

Figure 3: Spearman correlation between human judgment and model prediction, across 200 verbs. Judgments were averaged across themes and recipients, hence the lower overall preferences for the double-object.

tures the full spectrum of human verb biases (Fig. 3). To do this, we measured the Spearman correlation between human judgments and model predictions across the 200 verbs, averaging over recipients and themes. We found that the transformer architectures are particularly sensitive to human-like verb biases, with the larger GPT2 model having the highest correlation ($r = 0.73$). The larger LSTM model had an even greater number of parameters but accounted for significantly less variance, suggesting that simply increasing model size may not be sufficient to learn verb-specific preferences (van Schijndel et al., 2019).

Next, we examined the extent to which each model qualitatively accounts for human sensitivity to argument length and definiteness (Fig. 2B), averaging across verbs. For all models except the n-gram model, we found significant effects of recipient length, recipient definiteness, and theme definiteness (see Table S5 in Appendices for details). Overall, however, the LSTM models were more sensitive to the effect of definiteness, showing the same additive effects as human speakers. Additionally, all models except BERT reflected the fact that ratings on the DO are highest when the recipient is labeled by a (definite) pronoun.

4.3 Probing internal representations

Having established key differences in the predictive accuracy of different models, we now investigate the internal representation of this knowledge. We hypothesized that sensitivity to verb bias requires the ability to integrate the verb's lexical embedding with the higher-level structure of the sentence. Thus, successful models should contain information about acceptability early in the sentence.

To focus on verb bias, we began with the subset of 1000 sentences with pronoun recipients. For the 4 auto-regressive models (two different sizes of LSTMs and GPT-2), we then extracted the hidden state after each word. To analyze how acceptability was represented throughout the sentence, we fit regularized linear regressions using the hidden state features as input and human judgments as output (see Appendix C for more details). We then compared these predictions at three key points in the sentence: after the verb, after the first argument, and after the second argument.

Upon seeing the verb, human preferences for the DO were already decodable from the GPT2 models' features with higher precision than from the LSTMs' (see Fig. 4), reflecting richer lexical representations. At the same time, predictive accuracy increased for all models after the first argument, reflecting additional cues from the word sequence. For example, the model may represent that a pronoun recipient appearing after a PO-biased verb is likely to be less acceptable (e.g. **Alice said him...*). Finally, we observed that all of the models lost information about construction preference near the end of the sentence.

How does the representation of verb bias change as a function of depth in the best-fitting GPT2-large architecture? Recent analyses have found that the ability to decode syntactic information peaks near the middle layers of transformers (Tenney et al., 2019; Hewitt and Manning, 2019). In the previous analysis, we took the single layer that maximized explained variance; here, we repeat this analysis across all layers (Fig. 5). Immediately after observing the verb, decodability of DO preferences is

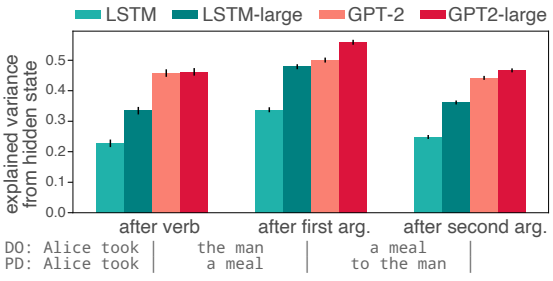

Figure 4: Variance predicted by hidden states throughout the sentence. Error bars show cross-validated SEM.

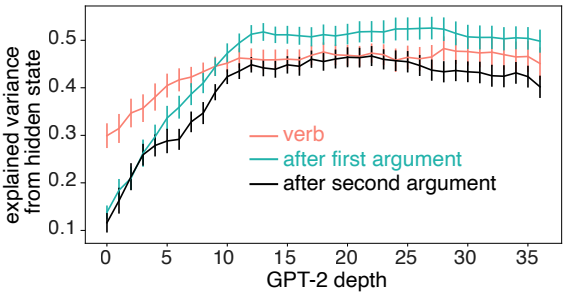

Figure 5: Decodability at each layer of GPT2-large.

already high in the earliest layers, suggesting that this information is directly available from the verb's lexical embedding. Later in the sentence, however, DO preferences are no longer decodable at lower layers; instead, it has shifted to intermediate layers, suggesting increasing reliance on context and higher-order structure.

5 Analysis on natural corpus

While there are many advantages of human judgment datasets like DAIS — including the ability to include a wider range of infrequently observed verbs and to control for potential confounds — there are also distinct advantages of corpus data. We thus conducted a further evaluation using DO and PO utterances extracted from the Switchboard corpus by Bresnan et al. (2007). Instead of testing how well each model was able to predict continuous judgments, we now ask how well each model is able to categorically predict whether the DO vs. PO was naturally produced by a corpus speaker.

For each DO or PO utterance in the corpus, we used the extracted verb, theme, and recipient to generate the alternating sentence, pairing the attested example with its hypothetical alternation. Subjects were chosen from a list of names, as in the DAIS dataset. After removing incoherent sentences, we obtained 2,206 pairs of sentences in total.

For each model, we calculated the likelihood ratio of the PO vs. DO construction for all corpus examples. Following Bresnan et al. (2007), we then constructed a classifier to predict which construction was actually produced by fitting a decision threshold for likelihood ratios. The accuracy achieved by each model is shown in Table 2, reproducing roughly the same ranking that we observed for the DAIS dataset. Critically, the GPT2 models achieved comparable accuracy to the 92% previously reported by Bresnan et al. (2007) using a logistic regression on 14 hand-annotated binary features (e.g. animacy, accessibility, definiteness).

GPT2 (large)	GPT2	BERT	LSTM (large)	LSTM	Ngram
93.51	93.47	91.29	88.21	81.13	80.05

Table 2: Classification accuracy on Switchboard

An important difference between Switchboard datives and the DAIS dataset is the set of verbs represented. Switchboard only contains 37 distinct verbs compared to our 200, and is heavily skewed by frequency ('give' accounts for 42% of examples). Additionally, while we intentionally included 100 verbs traditionally considered "non-alternating," the 37 verbs in Switchboard are skewed toward "alternating." Of the 27 of these appearing in (Levin, 1993), all were classified as "alternating," and all but one of the 37 appeared in the double-object construction at least once in the dataset. These features may help account for why Switchboard was unable to distinguish between our GPT2 models: it may be an easier task than DAIS.

6 Conclusions

In natural languages, speakers routinely select one alternative over others to express their intended message. These choices are sensitive to many interacting factors, including the choice of the main verb and the length and definiteness of arguments. Our new dataset, DAIS, not only offers a higher-resolution window into the richness of human preferences, it also provides a newly powerful benchmark for evaluating and understanding the corresponding sensitivity of language models. We found that transformer architectures corresponded especially well with human verb bias judgments.

Further work is needed to more precisely determine the source of the architectural differences we observed. One possibility is that the transformer's self-attention mechanism and layer-wise organization improves its ability to represent lexically-specific structures. However, it is also possible that differences are attributable to training data. Another line of future research is to compare the incremental predictions of neural models to finer-grained eye-tracking evidence during sentence processing of double-object sentences (e.g. Filik et al., 2004). As neural language models become more complex, subtler phenomena like verb bias may yield new insights into how lexical and grammatical representations are jointly learned and successfully integrated for language understanding.

Acknowledgements

This work was supported by a Ma Huateng Data Driven Science SML award to AEG, NSF grant #1718550 to TLG, and NSF grant #1911835 to RDH. This work was developed while TY was visiting Princeton, supported by the Ito foundation USA-FUTI scholarship. RDH and TY contributed equally and share joint first authorship. We are grateful to Tom Wasow, Joan Bresnan, and Tatiana Nikitina for providing corpus materials and thoughtful comments.

References

Ben Ambridge, Libby Barak, Elizabeth Wonnacott, Colin Bannard, and Giovanni Sala. 2018. Effects of both preemption and entrenchment in the retreat from verb overgeneralization errors: Four reanalyses, an extended replication, and a meta-analytic synthesis. *Collabra: Psychology*, 4(1).

Sharon Lee Armstrong, Lila R Gleitman, and Henry Gleitman. 1983. What some concepts might not be. *Cognition*, 13(3):263–308.

Jennifer E Arnold, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1):28–55.

Libby Barak, Afsaneh Fazly, and Suzanne Stevenson. 2014. Learning verb classes in an incremental model. In *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics*, pages 37–45.

Joan Bresnan. 2007. Is syntactic knowledge probabilistic? Experiments with the english dative alternation. In Sam Featherston and Wolfgang Sternefeld, editors, *Roots: Linguistics in search of its evidential base*, pages 75–96. Mouton de Gruyter Berlin.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R Harald Baayen. 2007. Predicting the dative alternation. In I. Kraemer G. Boume and J. Zwarts, editors, *Cognitive foundations of interpretation*, pages 69–94. Royal Netherlands Academy of Science, Amsterdam.

Joshua R De Leeuw. 2015. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, 47(1):1–12.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Ruth Filik, Kevin B Paterson, and Simon P Liversedge. 2004. Processing doubly quantified sentences: Evidence from eye movements. *Psychonomic Bulletin & Review*, 11(5):953–959.

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

Edward Gibson and Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1-2):88–124.

Adele E Goldberg. 2019. *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Jess Gropen, Steven Pinker, Michelle Hollander, Richard Goldberg, and Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in english. *Language*, pages 203–257.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL-HLT*, page 1195–1205.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of ACL*, pages 690–696.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of NAACL*, pages 4129–4138.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R Bowman. 2018. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 287–297.

M Alex Kelly, Yang Xu, Jesús Calvillo, and David Reitter. 2020. Which sentence embeddings and which layers encode syntactic structure? In *Proceedings of the 42nd Conference of the Cognitive Science Society*.

Steven Langsford, Amy Perfors, Andrew T Hendrickson, Lauren A Kennedy, and Danielle J Navarro. 2018. Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: a journal of general linguistics*, 3(1).

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1).

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*.

Paul Marty, Emmanuel Chemla, and Jon Sprouse. 2019. The effect of three basic task features on the sensitivity of acceptability judgment tasks. *Manuscript. https://ling. auf. net/lingbuzz/004588*.

Richard Thomas Oehrle. 1976. *The grammatical status of the English dative alternation*. Ph.D. thesis, Mass. Cambridge.

Amy Perfors, Joshua B Tenenbaum, and Elizabeth Wonnacott. 2010. Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of child language*, 37(3):607–642.

Robert J Podesva and Devyani Sharma. 2014. *Research methods in linguistics*. Cambridge University Press.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Rachel A Ryskin, Zhenghan Qi, Melissa C Duff, and Sarah Brown-Schmidt. 2017. Verb biases are shaped through lifelong learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(5):781.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. In *Proceedings of EMNLP*, page 5831–5837.

Martin Schrimpf, Idan A Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy G Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2020. Artificial neural networks accurately predict language processing in the brain. *BioRxiv*.

Carson T Schütze. 2011. Linguistic evidence and grammatical theory. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(2):206–221.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of ACL*, page 4593–4601.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov Random Field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation (NeuralGen)*, page 30–36.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Thomas Wasow. 2002. *Postverbal behavior*. CSLI Stanford, CA.

Charles Yang. 2008. The great number crunch. *Journal of Linguistics*, 44(1):205–228.

Appendix A: Data collection details

There are many possible ways of empirically eliciting acceptability judgements (Marty et al., 2019; Podesva and Sharma, 2014; Langsford et al., 2018). We chose to present pairs of sentences together with a continuous slider to maximize our power to detect gradient preferences. We generated a sentence pair for each verb-theme item by randomly selecting a subject from a list of 8 names (e.g. Juan, Alice), and selecting recipients from a short list corresponding to the given condition (e.g. "him," "her," or "them" for the pronoun condition; "the man," "the woman," "the team" for the short definite condition, etc.) See Table S1 for examples. We implemented our study using jsPsych (De Leeuw, 2015) and paid participants a $1.00 base pay in addition to an additional $1.00 completion bonus.

To ensure data quality, we excluded participants who failed an initial comprehension quiz or either of two attention checks where one of the sentences in the pair was randomly scrambled:

(4) a. The man ate a slice of cake.
b. The man cake of slice ate a.

We also excluded individual trials with response times of < 3 seconds, and all trials from participants who responded this quickly for more than a quarter of their responses, since it was not possible to read the sentences in that time. Due to these exclusions, as well as generic participant dropout on Mechanical Turk, not all sentences received the same number of judgements, but we ensured that at least 5 judgements were collected for each sentence pair.

Appendix B: Regression specifications

To evaluate the binary effect of alternating vs. non-alternating verbs in Section 4.1, we constructed a mixed-effects model predicting human preferences including a dummy-coded fixed effect for the "alternating" vs. "non-alternating" classification from

Levin (1993). We also included random intercepts and slopes for each human participant.

To evaluate the effect of information structure in Section 4.1, our mixed-effects model included fixed effects for recipient length, recipient definiteness, and theme definiteness. We included random intercepts and effects of recipient length and definiteness for each participant and verb to control for clustered variance at these levels. See Fig. S1 for the full pattern of results, split by "alternating" and "non-alternating" verbs. Complete regression results are shown in Tables S3 and S4.

Appendix C: Analysis details

For each of three sentence positions of interest investigated in section 5 (after verb, after first argument, and after second argument), we fit a linear regression predicting human judgements from the hidden states. Because of the high dimensionality of these states, we used ridge regression to prevent overfitting[3]. The ridge regression regularization hyper parameter was optimized for each regression model through a log-scale grid search ($\alpha \in [10^0, 10^7]$) on a held-out validation set. As our evaluation metric, we computed R^2, or variance explained. Results were averaged across 10 runs of cross-validation, using random 80/20 splits (see Table S2 for best-performing hyperparameter configurations).

Because the predicted judgements were relative preferences between the two sentences, we concatenated the hidden states of the two sentences together as input. For the 2-layer LSTMs, we used the final hidden state. For the deeper GPT-2 architectures, which are known to represent different information at different layers, we did not know *a priori* which layer would be most appropriate. We thus conducted the regression analysis separately for each layer, and reported the highest performance that was achievable by the model across all layers. In other words, we computed the cross-validated mean performance for each layer and selected the best. This approach has also been used in other recent work (Schrimpf et al., 2020).

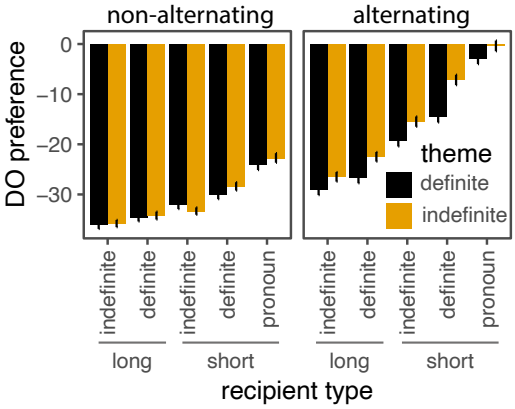

Figure S1: Full pattern of human recipient and theme effects for alternating and non-alternating verbs.

[3]We used the `sciki-learn` implementation.

DO sentence	**PO sentence**
Michael transported her the food	Michael transported the food to her
Bob recited the woman something	Bob recited something to the woman
Juan took a woman a gift	Juan took a gift to a woman
Alice supplied the man who was from work the news	Alice supplied the news to the man who was from work

Table S1: Example sentence pairs

	LSTM	LSTM-large	GPT2	GPT2-large
after verb	$1.3(\pm\ 0.2) \times 10^2$	$4.3(\pm\ 0.2) \times 10^2$	$1.0(\pm\ 0.1) \times 10^4$	$2.4(\pm\ 0.3) \times 10^4$
after 1st arg.	$1.1(\pm\ 0.2) \times 10^2$	$2.3(\pm\ 0.2) \times 10^2$	$6.3(\pm\ 0.6) \times 10^3$	$1.9(\pm\ 0.2) \times 10^4$
after 2nd arg.	$1.1(\pm\ 0.2) \times 10^2$	$1.8(\pm\ 0.2) \times 10^2$	$1.9(\pm\ 0.1) \times 10^3$	$3.6(\pm\ 0.4) \times 10^3$

Table S2: Regularization hyperparameter configuration for each model and task. SEM across cross-validation runs in parentheses.

term	estimate	t statistic	df	p value
(Intercept)	36.44	31.02	229.52	$< 1.0 \times 10^{-32}$
recipient length long vs pronoun	-16.27	-21.15	257.31	$< 1.0 \times 10^{-32}$
recipient length short vs pronoun	-8.00	-18.19	281.29	$< 1.0 \times 10^{-32}$
recipient definite vs. indefinite	-3.91	-14.41	194.96	$< 1.0 \times 10^{-32}$
theme definite vs indefinite	2.23	10.99	46616.18	$< 1.0 \times 10^{-32}$

Table S3: Fixed effect estimates for human mixed-effects regression, including random effects at the verb-level and participant level. Recipient length, recipient definiteness, and theme definiteness are dummy coded.

random group	term	estimate
participant	sd(Intercept)	9.18
participant	cor(Intercept, recipient length long vs pronoun)	-0.52
participant	cor(Intercept, recipient length short vs pronoun)	-0.45
participant	cor(Intercept, recipient definite vs. indefinite)	-0.76
participant	sd(recipient length long vs. pronoun)	8.96
participant	cor(recipient length long vs pronoun, short vs. pronoun)	0.84
participant	cor(recipient length long vs pronoun, definite vs indefinite)	0.90
participant	sd(recipient length short vs. pronoun)	6.81
participant	cor(recipient length short vs pronoun, definite vs indefinite)	0.91
participant	sd(recipient definite vs indefinite)	0.52
verb	sd(Intercept)	15.70
verb	cor(Intercept, recipient length long vs pronoun)	-0.93
verb	cor(Intercept, recipient length short vs pronoun)	-0.76
verb	cor(Intercept, recipient definite vs. indefinite)	-0.80
verb	sd(recipient length long vs pronoun)	9.22
verb	cor(recipient length long vs pronoun, short vs. pronoun)	0.92
verb	cor(recipient length long vs pronoun, definite vs indefinite)	0.76
verb	sd(recipient length short vs. pronoun)	3.48
verb	cor(recipient length short vs pronoun, definite vs indefinite)	0.66
verb	sd(recipient definite vs indefinite)	2.19
Residual	sd(observation)	22.25

Table S4: Random-effect estimates for mixed-effects regression on human judgments.

Figure S2: Histograms of individual slider responses for all 200 verbs. Verbs are ranked from lowest mean preference for DO to highest mean preference for DO. Verbs classified as "non-alternating" by Levin (1993) colored red, "alternating" colored blue.

model	regression term	estimate	t statistic	df	p value	sig. level
bert	(Intercept)	-2.83	-6.66	118.43	9.22e-10	***
bert	recipient length pronoun vs. long	-6.08	-12.75	99.34	1.23e-22	***
bert	recipient length pronoun vs. short	2.59	8.00	143.07	3.91e-13	***
bert	recipient definite vs. indefinite	-5.68	-18.98	99.08	9.10e-35	***
bert	theme definite vs. indefinite	4.00	19.99	2198.00	7.95e-82	***
gpt2	(Intercept)	1.01	5.59	121.11	1.43e-07	***
gpt2	recipient length pronoun vs. long	-6.43	-29.23	100.52	6.02e-51	***
gpt2	recipient length pronoun vs. short	-2.44	-17.44	202.93	3.09e-42	***
gpt2	recipient definite vs. indefinite	-0.25	-2.00	99.42	4.80e-02	*
gpt2	theme_ypeindef	0.96	11.13	2198.00	5.14e-28	***
gpt2-large	(Intercept)	0.20	1.09	116.31	2.80e-01	n.s.
gpt2-large	recipient length pronoun vs. long	-5.81	-27.91	99.00	1.02e-48	***
gpt2-large	recipient length pronoun vs. short	-1.78	-12.85	99.00	7.96e-23	***
gpt2-large	recipient definite vs. indefinite	-0.57	-4.80	99.00	5.65e-06	***
gpt2-large	theme definite vs. indefinite	1.44	17.00	2099.00	8.04e-61	***
lstm	(Intercept)	-1.85	-9.02	124.11	2.92e-15	***
lstm	recipient length pronoun vs. long	-2.80	-8.80	100.14	4.07e-14	***
lstm	recipient length pronoun vs. short	-0.87	-5.26	219.65	3.44e-07	***
lstm	recipient definite vs. indefinite	-1.33	-12.04	1464.63	7.00e-32	***
lstm	theme definite vs. indefinite	1.61	16.04	2297.00	6.16e-55	***
lstm-large	(Intercept)	-1.19	-3.05	136.46	2.74e-03	**
lstm-large	recipient length pronoun vs. long	-9.38	-20.77	105.00	5.73e-39	***
lstm-large	recipient length pronoun vs. short	-2.30	-6.98	411.84	1.16e-11	***
lstm-large	recipient definite vs. indefinite	-1.02	-3.73	100.60	3.16e-04	***
lstm-large	theme definite vs. indefinite	3.21	14.67	2198.00	1.47e-46	***
ngram	(Intercept)	1.27	13.27	124.45	1.39e-25	***
ngram	recipient length pronoun vs. long	-1.93	-19.59	107.84	2.60e-37	***
ngram	recipient length pronoun vs. short	-1.26	-12.86	107.83	1.68e-23	***
ngram	recipient definite vs. indefinite	-0.04	-0.72	98.99	4.72e-01	n.s.
ngram	theme definite vs. indefinite	0.87	16.59	2197.99	2.59e-58	***

Table S5: Mixed-effects regression results for each model, including random effects at the verb-level. Recipient length, recipient definiteness, and theme definiteness are dummy coded. *** denotes $p < 0.001$, ** denotes $p < 0.01$, * denotes $p < 0.05$, n.s. denotes 'not significant.'