A Discerning Several Thousand Judgments: GPT-3 Rates the Article + Adjective + Numeral + Noun Construction

Kyle Mahowald
mahowald@utexas.edu
Department of Linguistics
The University of Texas at Austin

Abstract

Knowledge of syntax includes knowledge of rare, idiosyncratic constructions. LLMs must overcome frequency biases in order to master such constructions. In this study, I prompt GPT-3 to give acceptability judgments on the English-language Article + Adjective + Numeral + Noun construction (e.g., "a lovely five days"). I validate the prompt using the CoLA corpus of acceptability judgments and then zero in on the AANN construction. I compare GPT-3's judgments to crowdsourced human judgments on a subset of sentences. GPT-3's judgments are broadly similar to human judgments and generally align with proposed constraints in the literature but, in some cases, GPT-3's judgments and human judgments diverge from the literature and from each other.

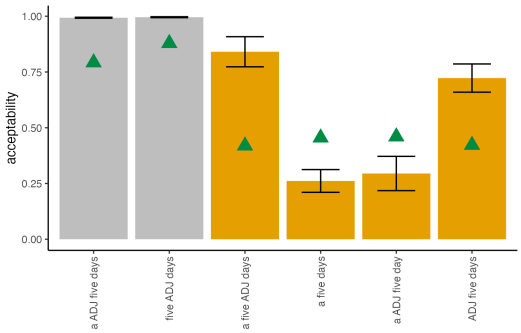

Figure 1: GPT-3 acceptability judgments (bars), compared to human ratings (green triangles) on a matched set of sentences. The comparison is between the AANN construction and the standard alternative, as well as 4 degenerate versions. Both humans and GPT-3 rank the AANN construction as being as acceptable as the standard, and all degenerate constructions are rated significantly lower.

1 Introduction

Consider the English Article + Adjective + Numeral + Noun (AANN) construction: " **a beautiful 228 pages** [iWeb]" or "The president has had **a terrible five weeks** [COCA]". Usually cardinal numbers precede the adjective ("five terrible weeks"), but here the adjective precedes the numeral. More strangely, the normally singular article "a" in this case is followed by a plural noun phrase.[1]

An eclectic dozen or so papers have been written on the construction, many focused on elucidating relevant semantic and syntactic constraints (Goldberg and Michaelis, 2017; Jackendoff, 1977; Dalrymple and King, 2019; Bylinina and Nouwen, 2018; Ionin and Matushansky, 2018, 2004; Solt, 2007; Keenan, 2013). The presence of the modifier is crucial: "a 228 pages" is unacceptable. The type of modifier is also crucial: "a pink 228 pages" is odd because color words are "stubbornly distributive" (Schwarzschild, 2011) and thus cannot refer to a set of items as a whole: the nominal phrase needs to function as a unit (Solt, 2007). These idiosyncratic constraints are typical of *constructions* (Goldberg, 2019).

Prior work on the AANN construction has focused on characterizing the semantic and syntactic constraints on the construction and proposing analyses in various frameworks. For instance, Solt (2007) focuses on how the construction coerces the phrase into a singular noun phrase; Keenan (2013) proposes treating it as akin to a partitive, and Dalrymple and King (2019) give an LFG analysis.

The same properties that make AANN interesting from the perspective of human language use—its low frequency but high sensitivity to constraints—also make it interesting from the perspective of LLMs and what they learn about linguistic structure. Indeed, much work on LLM syntactic competence has centered on ubiquitous abstract features of grammar like subject-verb number agreement (e.g., Linzen et al., 2016; Gulordava et al., 2018), part of speech (Tenney et al., 2019), and syntactic dependencies (e.g., Hewitt and Manning,

[1]Data and code: https://github.com/mahowak/aann-public/

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 265–273
May 2-6, 2023

2019). That said, there has also been a recent spate of research on construction-grammar-inspired approaches in NLP, including studies showing that LLMs have access to construction information (Tayyar Madabushi et al., 2020; Tseng et al., 2022; Weissweiler et al., 2022) and capture verb argument construction biases (Hawkins et al., 2020) as well as fine-grained lexical semantic information (Petersen and Potts, 2022). Moreover, sentences with similar constructions cluster in embedding space (Li et al., 2022).

These works are valuable because better understanding how LLMs handle constructions could help us better understand what LLMs learn about linguistic structure (Baroni, 2021; Linzen and Baroni, 2021) and could also inform us as to what can be learned about language from primary data (Warstadt and Bowman, 2022, 2020). In our case, for an LLM to get the AANN construction right, a number of statistical regularities must be eschewed: "a" cannot be treated as a singular marker since the noun is plural, the normal ordering of the number and adjective must be reversed, and normal verb number agreement rules must in some cases be suspended. Understanding whether these heuristics (which work well for the vast majority of text) can be overcome can guide us towards future work understanding *how* they are overcome.

Here, I ask what GPT-3 `text-davinci-002` (now often classed as an instance of GPT-3.5) learns about the AANN construction by testing its sensitivity to several constraints proposed in the literature. In doing so, I treat the LLM as a linguistic test subject (Linzen et al., 2016; Futrell et al., 2019; Wilcox et al., 2021; Warstadt et al., 2019; Ettinger, 2020). I use a custom prompt to elicit quantitative grammaticality judgments (Schütze, 2016; Gibson and Fedorenko, 2013) from GPT-3, and show that the prompt performs well on CoLA, a data set of binary acceptability judgments on a carefully constructed 10,657 English sentences (Warstadt et al., 2019). I then unleash it on the AANN construction and compare GPT-3 to human ratings.

2 Methods

Attaining acceptability judgments from language models is not straightforward. Simply comparing sentence probabilities is difficult because they are dependent on the individual lexical items, as well as sentence length (Warstadt et al., 2020). I rely on the prompting paradigm to elicit acceptability judgments—validating the measure first using a known data set of judgments from CoLA. The prompt was created by drawing on a combination of CoLA training sentences and handcrafted sentences and iteratively experimenting. Ultimately, a diverse 8 sentences were chosen, some of which intentionally have low lexical probability to ensure that the model does not call all low-probability strings ungrammatical. Each prompt example sentence appears along with a binary judgment: "good" or "bad". Then, GPT-3 is passed the prompt, along with the critical test sentence, and asked to generate one more token (always either "good" or "bad"). The probability of the generated word (whether "good" or "bad") is our numerical rating.

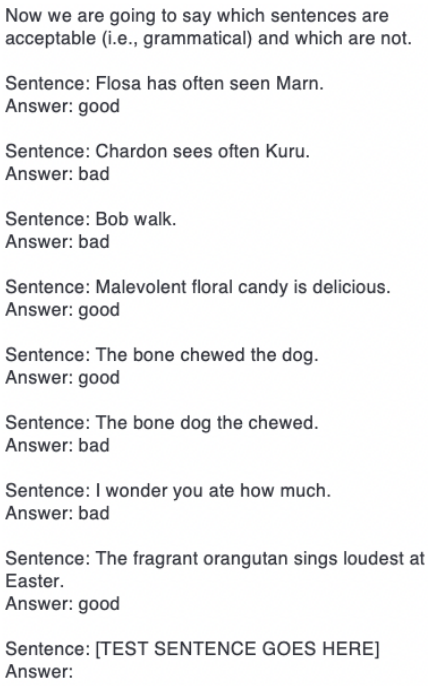

Figure 2: The prompt used for attaining grammaticality judgments. The target sentence is inserted, and then GPT-3 is asked to generate one token (overwhelmingly likely to be "good" or "bad"). That token's probability is taken as a rating.

To validate the measure, I tested the final prompt on the CoLA dev set. It attained accuracy of 84%, with a Matthew's correlation coefficient of 0.63. This is worse than, but in the ballpark of, human inter-annotator agreement on CoLA which is 86% and .697, respectively (Warstadt et al., 2019). It is also comparable to top performers on the GLUE leaderboard for the CoLA sub-task.

I used this technique to test the AANN con-

template	**temporal**: The family spent X in London; The diplomat worked X in Nairobi; The tourist stayed X in Papua New Guinea; **objects**: She bought X; They discovered X; Someone saw X; **human**: We served dinner to X; X greeted us at the door; We congratulated X; **art**: The newspaper reviewed X; I experienced X; Please enjoy X; **distance**: He drove X; Someone walked X; Someone traveled X; **unitlike**: Luis took in X; They consumed X; It lasted X
template for agreement task	**temporal**: X is/are just what you need; X is/are ideal **objects**: X is/are available; X make(s) a lovely gift; **human**: X regularly show(s) up at the door; X is/are here **art**: X was/were reviewed in the newspaper; X was/were enjoyed **distance**: X is/are a long way; X is/are not far; **unitlike** X was/were uncovered; X was/were make(s) an impression
adj	**ambig**: astonishing; incredible; impressive; disappointing; surprising; devastating; pathetic; remarkable; mediocre; unsatisfying; **qualitative**: lovely; beautiful; enchanting; soothing; charming; disgusting; uninviting; haunting; hideous; ugly; **quant**: mere; staggering; whopping; hefty; paltry; meager; extra; measly; substantial; record-setting; **stubborn**: large; big; small; round; tall; **color**: blue; green; red; yellow; orange; **human**: lucky; talented; graceful; fancy; friendly; collegial; hopeful; shy; bold; grinning
noun	**human**: soldiers; students; athletes; pianists; teammates; lawyers; doctors; actors; Americans; bankers; **objects**: desks; marbles; pencils; belts; forks; chairs; cans; bananas; apples; trays; **art**: movies; paintings; books; shows; operas; **temporal**: days; weeks; months; years; hours; **distance**: meters; feet; yards; blocks; steps; **unit_like**: pages; acts; paragraphs; awards; meals
num.	three; five; six; eight; ten; twenty; fifty; 500; 1000; 10,000; 21; 51; 512; 1,429; 21,234

Table 1: A superset of the items used, which were combined in various ways across experiments. In the templates, X is replaced by the AANN construction.

struction by templatically constructing sentences in which I parametrically vary the main sentence template, adjective, numeral, and nominal, from the superset shown in Table 1. Templates were designed to work with the key manipulations. Certain nouns work with some templates and not others, to ensure template/ noun pairs are always plausible.

Adjectives in the AANN construction behave differently depending on whether they are quantitative (i.e. modify the numeral as in "a mere 5 days"), qualitative (e.g., modify the noun as in "a beautiful five days), or are ambiguous between the two (e.g., "an astonishing five days" which leaves it unclear whether the number of days is astonishing, or the days themselves). It has been claimed (e.g., Dalrymple and King, 2019; Solt, 2007; Keenan, 2013) that quantitative and ambiguous adjectives are typically more acceptable than qualitative ones in AANN—although there are specific instances where qualitative adjectives are acceptable. I also consider "stubbornly distributive" (Schwarzschild, 2011) adjectives (e.g., "large" or "blue"), which "stubbornly" refer to individuals even when applied to a group. For instance, "The chairs are large." can refer only to the individual chairs being large, not to the collective group of chairs being large. It's claimed (Bylinina and Nouwen, 2018; Ionin and Matushansky, 2018; Keenan, 2013) that this same property makes a phrase like "a large five trees" less acceptable, compared to something like "a beautiful five trees" (in which it is possible for "beautiful" to refer to not just the individual trees but to the collection of trees).

Solt (2007) and others observe measure nouns work best in the AANN construction, but that other nouns can be okay as long as they can be treated as a single unit. I sampled nouns from 6 categories, as shown in Table 1. Also as shown in that table, I sampled numerals of various kinds but focused on "three" and "five" for the human experiments and most analyses. See Appendix B for a discussion of sensitivity to the numeral.

From these templates and candidate words, I generate semantically plausible sentences (meaning that, throughout the experiments, I only use human-appropriate sentence templates with human nouns and object-appropriate ones with object nouns). Depending on the specific question in each experiment, I run controlled subsets of these sentences (or their degenerate variants as in Experiment 1) through GPT-3 to obtain acceptability judgments, getting the probability of "good" or "bad" as the next word in the continuation.

I also use Amazon's Mechanical Turk to obtain human judgments on a subset of the test sentences. I asked raters to rate sentences on a rating bar scale from 1-10. For Experiments 1 and 3, each rater rated 3 critical sentences. They rated 18 critical sentences in Experiment 2 since there were many more conditions to test in that experiment.

To guard against raters becoming inured to the construction, half or more items were fillers not involving AANN. To maintain similar calibration between humans and GPT-3, these fillers always included all example judgments used in the GPT-3 prompt. I excluded participants who did the survey more than once, who did not rate the good filler items at least 1 point higher than the bad, or who did not have a US IP address. I obtained annotations for only a sample of sentences rated by GPT-3. When applicable, analyses focus on the union of sentences rated by both humans and GPT-3.

3 Exp. 1: AANN Fundamentals

First, I tested the basics of the AANN construction as laid out in the literature by having GPT-3 and human raters rate the AANN construction (e.g., "a beautiful five days") vs. the default ("five beauti-

ful days") vs. a select four degenerate conditions: one with the order of the numeral and adjective switched, one with no modifier, one with a singular noun, and one with no article "a". To generate examples in each condition, I crossed 3 temporal nouns (days, weeks, months), with a low numeral (three), one of 14 appropriate adjectives, and one of 3 templates. For the resulting sentences, I attained human ratings from MTurk (after exclusions, 126 raters rating 3 sentences each, for 378 total ratings) and GPT-3 and focus on that subset for analysis.

Figure 1 shows results for this experiment. Although GPT-3 uses a wider range of the scale, both rate the AANN construction as just as good as the default and give lower ratings to the 4 degenerate versions. Humans rate the 4 degenerate constructions as about equally bad (all between .46 and .53), whereas GPT-3 rates the versions with swapped adjective/numeral order and a missing article as significantly better than the versions with a missing plural and a missing modifier. Running a mixed effect regression (Bates et al., 2015; Barr et al., 2013) comparing the AANN construction to the "default" construction and the degenerate alternatives (treating the default construction as the baseline, with random intercepts for adjective class, adjective, and template; and, for humans, for rater), both GPT-3 and humans show no significant difference in rating between AANN and default (both $p > .05$), but do show a significant difference between AANN and all 4 degenerate conditions (all $ps < .0001$). See Appendix C for regression details. Overall, I conclude that humans and GPT-3 "get" the AANN construction, even though they differ in the relative ratings of the bad variants.

4 Exp. 2: Adjectives and nouns

In this experiment, I focus on only the AANN construction and parametrically vary the kinds of adjectives (quantitative, ambiguously quantitative/qualitative, qualitative, human-referring, color adjective, stubbornly distributive; see Table 1 for examples) and kinds of nouns in the sentences (art, distance, human nouns, object nouns, temporal nouns, unit-like nouns; again see Table 1). This process produced a carefully controlled 12,960 unique sentences, from which we sampled a random subset for human ratings. After exclusions, there were 190 raters left who each rated 18 sentences, giving us ratings for 3,420 sentences.

I test whether GPT-3 and human raters agree

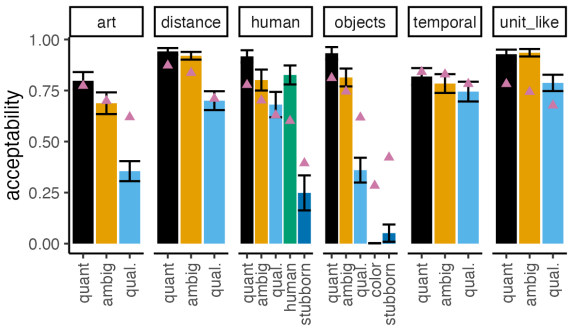

Figure 3: GPT-3 acceptability scores broken down by adjective type (x-axis), and noun type (on facets). Human ratings are pink triangles.

with the attested claims that (a) more measure-like nouns (e.g., temporal nouns, distance nouns, and unit-like nouns) are more acceptable in AANN, (b) qualitative adjectives are acceptable in only some AANN cases, and (c) stubbornly distributive adjectives (including color words) are not acceptable in the AANN construction. To assess significance, I predicted acceptability separately for humans and GPT-3 in a mixed effect regression, using adjective class, noun class, and their interaction as predictors and with random effects for adjective, noun, numeral, and template. I treated qualitative adjectives with temporal nouns as the baseline.

For both humans and GPT-3, there are significant differences in how nouns interact with adjectives (Figure 3). For temporal, distance, and unit-like nouns, all adjective types show high acceptability (although the qualitative adjectives are rated lowest). For art and object nouns, qualitative nouns score significantly worse ($p < .01$ for both humans and GPT-3) than ambiguous or quantitative nouns (an observation consistent with the literature). As predicted, colors and other stubbornly distributive adjectives (which can be tested only for humans and objects) show the lowest acceptability (significantly lower, $p < .01$, compared to qualitative adjectives, for both human and GPT-3 ratings). See Appendix D for regression details.

5 Exp. 3: Adjective Order

It is claimed that, in AANN, qualitative adjectives must appear before quantitative ones in order to be acceptable (Solt, 2007): "The family spent a beautiful mere five days in London." is preferred over "The family spent a mere beautiful five days in London.". To compare whether there is an effect of adjective ordering (putting the qualitative adjective

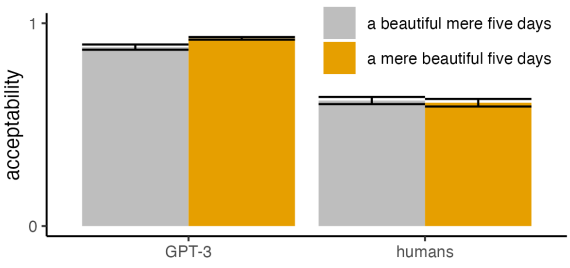

Figure 4: GPT-3 and human preference for adjective order (quantitative before qualitative; qualitative before quantitative).

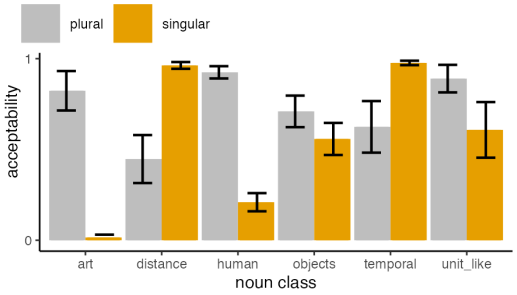

Figure 5: Mean GPT-3 acceptability ratings in the AANN construction for plural and singular verb agreement, as a function of noun class.

before the quantitative one or vice versa), I ran an experiment crossing 3 templates; 5 adjectives (astonishing, impressive, beautiful, hideous, ugly); the noun "days"; and the numeral "three" or "five". I ran each sentence in two conditions (quantitative adjective first or qualitative adjective first). For instance, I compared: "The family spent a beautiful mere five days in London." to "The family spent a mere beautiful five days in London." This left 60 sentences total, rated by 99 raters (each sentence rated between 18 and 36 times; 1,782 ratings total).

GPT-3 significantly prefers the order *dispreferred* in the literature (quantitative first, as in "a mere beautiful five days"; $\beta = .04, p < .01$). Humans (n=99) showed no clear preference according to the model, under the same analysis (but with a random intercept for rater): $\beta = -.513, p > .05$. Thus, the attested claim does not replicate. But the ratings for these sentences are relatively low overall, so we should remain open to the possibility that there are better examples of the double-adjective constructions than the ones tested. See Appendix E for regression details.

6 Exp. 4: Verb Agreement

The AANN construction also challenges number agreement. AANN subjects sometimes take singular verbs (when the noun phrase would be singular anyway, as in "A mere fifty cents for a cup of coffee sounds/*sound reasonable to me!"); sometimes plural ("A delicious four courses *was/were served in the main dining room."), and sometimes either ("A healthy two runs weekly was/were prescribed by the doctor." See Keenan (2013).

I tested agreement by comparing phrases which differed only in the verb number (e.g., comparing "A beautiful five days is..." vs. "A beautiful five days are...."). Sampling a subset of noun classes (art, distance, human nouns, objects, unit-like nouns, and temporal nouns), I generated a total of 280 sentences (each appearing with a singular or plural verb for a total of 560 sentences) and attained judgments from GPT-3. As shown in Figure 5, these results replicated in detail several attested judgments in the literature: art nouns, human nouns, unit-like nouns, and object nouns all prefer the plural for the AANN construction (art and humans almost categorically so). Distance nouns and temporal nouns prefer the singular (although temporal nouns can also take the plural). See Appendix F for details.

7 Conclusion

This work shows that GPT-3 can recognize and use the *form* of the AANN construction in a relatively (but not perfectly) human-like way, matching judgments across a variety of conditions, which is not the same thing as showing that it understands the meaning or *function* of the construction (Mahowald et al., 2023). In future work, we should study not just the form of the construction but its *construal* (Trott et al., 2020). This may be particularly relevant since Weissweiler et al. (2022), which studies the "Xer the Yer" construction (e.g., "the better the criticism, the better the science" [COCA]), show that LLMs can recognize the construction but fail at tests of understanding its meaning.

That said, GPT-3's performance on the AANN construction demands a significant amount of constructional knowledge and involves overriding major widespread "rules" of grammar (e.g., that the article *a* signals a singular noun). Future work exploring how LLMs override those heuristics, perhaps using causal intervention techniques (Ravfogel et al., 2021; Geiger et al., 2021) could illuminate their syntactic processing mechanisms.

8 Limitations

Many researchers have pointed out that the AANN construction is sensitive to context. For instance, Solt (2007) points out that "a hungry thirty hikers" may be acceptable in some sentences (namely ones where "hikers" is more easily construed as a single unit) than others. Because of the cost of running sentences through GPT-3 and on MTurk and the combinatoric nature of the construction, I could only run a constrained number of templates and did not consider larger context. Broader context may matter and so these results should be taken to apply to the particular contexts shown, which is why I show the exact templates used in the main text.

Another limitation is that the task of how to prompt GPT-3 for grammaticality judgments is not a settled question. Our focus in this paper was not on settling it, so I used one particular method for prompting GPT-3. While I explored the prompt space some, it is likely there are better prompts out there that would make GPT-3's performance on the task better. It's also possible that prompting GPT-3 for grammaticality judgments is not the best way to ascertain its knowledge of language and that a more naturalistic task would produce different results.

As has often been pointed out in the linguistics literature, naive human judges of out-of-context sentences may sometimes tap into different processes than they would when encountering language in the wild. Moreover, English is not a monolith and this construction's acceptability may vary across dialects of English. In a more detailed human study, it would be possible to tease apart effects of different dialects on ratings.

GPT-3 `text-davinci-002` is often categorized as GPT-3.5 because it is trained on more than just a word prediction task, and so we should not interpret its output as being purely reflective of what is learned by word prediction alone.

Finally, I note that I use templatically constructed sentences, which differ in important ways from naturalistic ones.

9 Acknowledgments

I acknowledge funding from NSF Grant 2104995. For helpful comments, I thank Robbie Kubala, Adele Goldberg, Gabriella Chronis, Katrin Erk, Alex Warstadt, Steve Wechsler, Liam Blything, attendees of the Princeton language group meeting, and an anonymous three reviewers.

References

Marco Baroni. 2021. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *arXiv preprint arXiv:2106.08694*.

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

Lisa Bylinina and Rick Nouwen. 2018. On "zero" and semantic plurality. *Glossa: a journal of general linguistics*, 3(1). Publisher: Open Library of Humanities.

Mary Dalrymple and Tracy Holloway King. 2019. An amazing four doctoral dissertations. *Argumentum*, 15(2019). Publisher: Debreceni Egyetemi Kiado.

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.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.

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.

Adele E Goldberg and Laura A Michaelis. 2017. One among many: Anaphoric one and its relationship with numeral one. *Cognitive science*, 41:233–258.

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.

Robert Hawkins, Takateru Yamakoshi, Thomas Griffiths, and Adele Goldberg. 2020. Investigating representations of verb bias in neural language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4653–4663, Online. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. 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 4129–4138. Association for Computational Linguistics. Event-place: Minneapolis, Minnesota.

Tania Ionin and Ora Matushansky. 2004. A singular plural. In *Proceedings of WCCFL*, volume 23, pages 399–412.

Tania Ionin and Ora Matushansky. 2018. *Cardinals: The syntax and semantics of cardinal-containing expressions*, volume 79. MIT Press.

Ray Jackendoff. 1977. X syntax: A study of phrase structure. *Linguistic Inquiry Monographs Cambridge, Mass*, pages 1–249.

Caitlin Keenan. 2013. "A pleasant three days in Philadelphia": Arguments for a pseudopartitive analysis. *University of Pennsylvania Working Papers in Linguistics*, 19(1):11.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423, Dublin, Ireland. Association for Computational Linguistics.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.

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. Publisher: MIT Press.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Erika Petersen and Christopher Potts. 2022. Lexical semantics with large language models: A case study of English *break*. Ms., Stanford University.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209, Online. Association for Computational Linguistics.

Carson Schütze. 2016. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Language Science Press.

Roger Schwarzschild. 2011. Stubborn distributivity, multiparticipant nouns and the count/mass distinction. In *Proceedings of NELS*, volume 39, pages 661–678. Graduate Linguistics Students Association, University of Massachusetts. Issue: 2.

Stephanie Solt. 2007. Two types of modified cardinals. In *International Conference on Adjectives. Lille*.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Sean Trott, Tiago Timponi Torrent, Nancy Chang, and Nathan Schneider. 2020. (Re)construing meaning in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5170–5184, Online. Association for Computational Linguistics.

Yu-Hsiang Tseng, Cing-Fang Shih, Pin-Er Chen, Hsin-Yu Chou, Mao-Chang Ku, and Shu-Kai Hsieh. 2022. CxLM: A construction and context-aware language model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6361–6369, Marseille, France. European Language Resources Association.

Alex Warstadt and Samuel R. Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data? In *Proceedings of the 42th Annual Meeting of the Cognitive Science Society - Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020, virtual, July 29 - August 1, 2020*. cognitivesciencesociety.org.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*, pages 17–60. CRC Press.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: A benchmark of linguistic minimal pairs for English. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 409–410, New York, New York. Association for Computational Linguistics.

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

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? probing pretrained language models for the english comparative correlative. *arXiv preprint arXiv:2210.13181*.

Ethan Wilcox, Pranali Vani, and Roger Levy. 2021. A targeted assessment of incremental processing in neural language models and humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 939–952, Online. Association for Computational Linguistics.

A Frequency of AANN

I sampled the AANN construction on SketchEngine for the English Web 2020 corpus with the following prompts: `[lemma="a[n]*"] [tag="JJ.*"] [tag="CD.*"] [tag="NNS"]` for AANN and `[tag="CD.*"] [tag="JJ.*"] [tag="NNS"]` for the vanilla construction. The AANN showed up 23.62 per million tokens, compared to 457.22 for the vanilla construction. Of the AANN construction examples, the vast majority contain quantitative adjectives (e.g., *mere*, *staggering*, etc.) and measurement nouns. Of a sample of 200 AANN constructions that I manually inspected on SketchEngine, none contained a qualitative adjective.

B Numerals

I sampled numerals from round low numbers ("three", "five", "six", "eight", "ten"), medium numbers ("twenty", "fifty"), high numbers (500, 1000, 10,000), non-rounded medium numbers (21 and 51), and non-rounded high numbers (1,429 and 21,234). I focused on round low numbers ("three" and "five") for most analyses.

For the overall analysis (across all adjectives and nouns in the main experiment), I get the below AANN ratings from GPT-3.

numclass	**example**	**avg. AANN score**
num-high	500	0.68
num-high_odd	1,429	0.93
num-low	three	0.80
num-med	twenty	0.70
num-med_odd	21	0.73

Low numbers like "three" and "five" are rated higher somewhat than other numbers, with the exception of high, non-round numbers (e.g., 1,429). These are rated unusually highly. Because of this anomalous behavior, I focus mostly on the low numerals for the analysis and did not attain human ratings for these other numerals. It remains an open question how humans would rate a sentence like "We spent a beautiful 1,429 days in London."

C AANN vs. default vs. degenerate regression

I run the following regression using the R `lme4` (Bates et al., 2015) package.

```
lmer(value ~ construction +
             (1|adjclass) +
             (1|adj) +
             (1 | temp))
```

where I treat the construction as a predictor (with the AANN constructor as a default) and adjclass, adjective, template as random effects. (Other random effects were removed to help the model converge, by iteratively removing ones with the smallest variance.)

For humans, there is also a random intercept for worker.

coef.	β_{gpt3}	Sig	β_{human}	Sig
(Intercept)	0.99	*	.81	*
five ADJ days	0.00		.06	
a five ADJ days	-0.22	*	-.39	*
a five days	-0.87	*	-.35	*
a ADJ five day	-0.72	*	-.35	*
ADJ five days	-0.20	*	-.38	*

Table 2: Fixed effect coefficients for GPT-3 comparing across constructions on the subset of sentences also run with human annotators

D Regression for adjective x noun sub-experiment with GPT-3 and humans

To assess significance, I predicted acceptability separately for humans and GPT-3 in a mixed effect regression, using adjective class, noun class, and their interaction as predictors and with random effects for adjective, noun, numeral, and template. I treated qualitative adjectives with temporal nouns as the baseline. For the human regression, it was identical except I included a random intercept for each worker, with a random slope for adjective

class and noun class (but not their interaction, due to convergence issues).

```
lmer(rating ~ adjclass *
       nounclass +
       (1|adj) +
       (1|noun) +
       (1|num) +
       (1 |template))
```

	beta	t-value	p<.05
(Intercept)	0.80	12.09	*
adj-quant	0.02	0.27	
adj-stubborn	-0.62	-10.27	*
adj-ambig	-0.02	-0.32	
noun-unit_like	0.14	3.49	*
noun-objects	0.02	0.58	
noun-human	0.03	0.76	
noun-distance	0.14	3.40	*
noun-art	-0.10	-2.33	*
adj-quant:noun-unit_like	-0.02	-0.80	
adj-quant:noun-objects	0.09	2.63	*
adj-stubborn:noun-objects	-0.17	-3.64	*
adj-quant:noun-human	0.08	2.24	*
adj-quant:noun-distance	-0.01	-0.28	
adj-quant:noun-art	0.07	2.35	*

Table 3: Fixed effect coefficients for GPT-3 comparing the adjective class x noun class manipulation.

	beta	t-value	p<.05
(Intercept)	0.73	16.84	*
adj-quant	0.10	2.63	*
adj-stubborn	-0.25	-5.58	*
adj-ambig	0.09	2.65	*
noun-unit_like	-0.08	-3.11	*
noun-objects	-0.09	-3.38	*
noun-human	-0.12	-4.25	*
noun-distance	0.00	0.17	
noun-art	-0.12	-4.28	*
adj-quant:noun-unit_like	0.02	0.72	
adj-quant:noun-objects	0.06	2.04	*
adj-stubborn:noun-objects	-0.04	-0.91	
adj-quant:noun-human	0.05	1.62	
adj-quant:noun-distance	0.03	1.27	
adj-quant:noun-art	0.04	1.51	

Table 4: Fixed effect coefficients for human annotators comparing the adjective class x noun class manipulation.

E Regression for adjective ordering

I ran a mixed effect linear regression predicting the GPT-3 score from the condition (quantitative-first vs. qualitative-first), with random intercepts for adjective, numeral, and template.

```
l = lmer(value ~ cond +
        (1|adj)  +
       (1|num) +
       (1 |template)
```

	beta	t-value	p<.05
(Intercept)	0.57	4.88	*
singular	0.35	3.14	*
noun-unit_like	0.27	3.37	*
noun-objects	0.17	2.50	*
noun-human	0.29	4.27	*
noun-distance	-0.18	-2.25	*
noun-art	0.20	2.52	*
singular:noun-unit_like	-0.63	-6.40	*
singular:noun-objects	-0.50	-6.12	*
singular:noun-human	-1.07	-12.96	*
singular:noun-distance	0.16	1.65	
singular:noun-art	-1.16	-11.71	*

Table 5: Fixed effect coefficients and significance values for an experiment comparing whether nouns in the AANN construction prefer singular or plural verbs.

F Agreement

I ran a regression predicting the rating based on the nounclass, and its interaction with whether there was singular plural. I included random intercepts for noun, adjective, and template, with a random slope for whether the verb was singular or plural on the noun factor.

```
lmer(rating ~ singplur *
      nounclass +
      (1 + singplur|noun) +
      (1|adj) +
      (1|template))
```

Results appear in Table 5, where the baseline values are temporal nouns in the plural (e.g., "days"). Singular verbs are preferred overall, relative to plurals for the temproal nouns (main effect of "singular"). There are various main effects of nouns, but the critical effects here are interactions. There are significant effects such that, *relative to temporal nouns* (e.g., "days"), unit-like nouns are less likely to prefer singular agreement, object nouns are less likely to prefer singular agreement, human nouns are less likely to prefer singular agreement, and art nouns are less likely to prefer singular agreement. Distance nouns are more likely than temporal nouns to prefer singular agreement (but not significantly so).