A Targeted Assessment of Incremental Processing in Neural Language Models and Humans

Ethan Gotlieb Wilcox
Harvard University
Department of Linguistics
wilcoxeg@g.harvard.edu

Pranali Vani
MIT
Brain and Cognitive Science
pvani@mit.edu

Roger P. Levy
MIT
Brain and Cognitive Science
rplevy@mit.edu

Abstract

We present a targeted, scaled-up comparison of incremental processing in humans and neural language models by collecting by-word reaction time data for sixteen different syntactic test suites across a range of structural phenomena. Human reaction time data comes from a novel online experimental paradigm called the *Interpolated Maze* task. We compare human reaction times to by-word probabilities for four contemporary language models, with different architectures and trained on a range of data set sizes. We find that across many phenomena, both humans and language models show increased processing difficulty in ungrammatical sentence regions with human and model 'accuracy' scores (à la Marvin and Linzen (2018)) about equal. However, although language model outputs match humans in direction, we show that models systematically under-predict the difference in magnitude of incremental processing difficulty between grammatical and ungrammatical sentences. Specifically, when models encounter syntactic violations they fail to accurately predict the longer reaction times observed in the human data. These results call into question whether contemporary language models are approaching human-like performance for sensitivity to syntactic violations.

1 Introduction

A substantial body of work has investigated contemporary language models (LMs) by assessing whether their behavior is consistent with the rules of syntax (Hu et al., 2020; Marvin and Linzen, 2018; Warstadt et al., 2020).[1] Among other structures, these studies have investigated agreement (Linzen et al., 2016; Gulordava et al., 2018) long distance dependencies (Wilcox et al., 2018), pronominal and particle licensing (Jumelet and Hupkes, 2018; Futrell et al., 2019), and expectations for phrase-level constituents (Futrell et al., 2018). Many of the studies which report aggregate behavior across a broad number of phenomena focus on accuracy scores, or the proportion of time LMs or human subjects in an online experiment prefer a grammatical variant in matching grammatical / ungrammatical sentence pairs. While these investigations provide much insight, they collapse a crucial dimension of comparison, namely the difference in magnitude between the grammatical and ungrammatical conditions. As long as the direction of their predictions are the same, an LM which finds grammatical conditions only marginally worse than their corresponding ungrammatical counterpart will receive the same score as a model that displays large differences between the two conditions.

At the same time, a related line of work has investigated the quantitative relationship between incremental predictions of language models and human reaction times (Hale, 2001; Levy, 2008). Smith and Levy (2013) found that this relationship is log-linear across multiple orders of magnitude for 3-gram models, and recent investigations have shown that this holds for contemporary neural network models as well (Wilcox et al., 2020; Goodkind and Bicknell, 2018). So far, this work has largely focused on the aggregate relationship, instead of isolating individual phenomena in targeted testing environments.

We combine these two approaches with a targeted assessment of incremental processing in neural language models and humans. We collect incremental processing data on a series of sixteen test suites, adapted from Hu et al. (2020), each of which targets a different syntactic phenomenon. For LM incremental processing data, we collect

[1] Data and code for this paper can be found online at https://github.com/wilcoxeg/targeted-assessment-imaze

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing, pages 939–952
August 1–6, 2021. ©2021 Association for Computational Linguistics

Test Suite Name	Tag	Example
Wh-Cleft Structures	Cleft	What she **did/spied** was **see the giraffe/the giraffe**
Filler-Gap Dependency, Subject Gap	FGD-subj	I know **who/that ___/my mother** sent the present to Taylor.
Filler-Gap Dependency, Object Gap	FGD-obj	I know **who/that** my mother sent **___/the present** to Taylor.
Filler-Gap Dependency, PP Gap	FGD-pp	I know **who/that** my mother sent the present to **___/Taylor** last weekend.
Main Verb/Reduced RC Gardenpath	MVRR	The ship **∅/that was sunk/steered** in the storm carried treasure.
NPI Licensing, *any*, Subj RC Modifier	NPL-any-src	**No/The** senator that **no/the** journalist likes has gotten any votes.
NPI Licensing, *any*, Obj RC Modifier	NPL-any-orc	**No/The** senator that likes **no/the** journalists has gotten any votes.
NPI Licensing, *ever*, Subj RC Modifier	NPL-ever-src	**No/The** senator that **no/the** journalist likes has ever won.
NPI Licensing, *ever*, Obj RC Modifier	NPL-ever-orc	**No/The** senator that likes **no/the** journalists has ever won.
Subject-Verb Number Agr., Subj RC Modifier	SVNA-src	The **lawyer/lawyers** that helped the mayor **is/are** organized.
Subject-Verb Number Agr., Obj RC Modifier	SVNA-orc	The **lawyer/lawyers** that the mayor hired **is/are** very organized.
Subject-Verb Number Agr., PP Modifier	SVNA-pp	The **lawyer/lawyers** next to the mayor **is/are** very organized.
Reflexive Anaphora, Masc., Subj RC Modifier	RNA-m-src	The **dukes/duke** that hunted the rabbits saw **himself/themselves** in the mirror.
Reflexive Anaphora, Masc., Obj RC Modifier	RNA-m-orc	The **dukes/duke** that the knights distrust saw **himself/themselves** in the mirror.
Reflexive Anaphora, Fem., Subj RC Modifier	RNA-f-src	The **queens/queen** that hunted the rabbits saw **herself/themselves** in the mirror.
Reflexive Anaphora, Fem., Obj RC Modifier	RNA-f-orc	The **queens/queen** that the knights distrust saw **herself/themselves** in the mirror.

Table 1: The sixteen test suites evaluated in this paper. Sentence regions which are manipulated to form the four conditions in each test suite are indicated with bold. Critical regions are underlined.

by-word probabilities for four contemporary neural network architectures. For human incremental processing data, we use by-word reaction times (RTs). We collect these by deploying a novel online measurement paradigm called the *Interpolated Maze*, which is based on the Maze task (Forster et al., 2009). In the Maze task, participants must read a sentence incrementally by selecting the correct word from two possible continuations, one of which is ungrammatical. The time it takes participants to select the correct choice has been shown to effectively capture incremental processing cost and can be deployed at scale (Boyce et al., 2020).

We deploy three analysis techniques to investigate how well models capture the human incremental processing data. First, we compute accuracy metrics (for LMs) and consistency scores (for humans) for each of our test suites, which correspond to the proportion of the time behavior is consistent with the relevant grammatical rules. We find that, for this analysis, humans and machine performance is about equal. Next, we compare the observed reaction-time slowdown between grammatical/ungrammatical conditions within a test suite to the slowdown predicted by each of our models. For this analysis we use the methodology developed by Van Schijndel and Linzen (2018), who use a *ms/bit* (milliseconds of reaction time per bit of surprisal) conversion metric derived from a fitted regression model to convert between the outputs of LMs and slowdowns in human reaction times. We find that models systematically under-predict the observed human data. In our third analysis, we train a linear regression models to predict reaction times from probabilities in non-critical sentence regions, and show that these models are relatively poor at predicting reaction times in critical sentence regions. That is, in areas of the sentence where human reaction time is influenced by grammatical violations, LM probabilities routinely under-predict human processing difficulty as measured by reaction time. Taken together, these results indicate that contemporary neural network languages models are systematically less sensitive to grammatical violations compared to humans.

2 Methods

We collect incremental processing data on a series of test suites, each of which targets an individual syntactic phenomenon. Composition of the test suites is described in Section 2.1. Methods used to collect incremental processing data are outlined in Section 2.2, for human reaction times. Section 2.3 describes the models tested. Linear Regression Models used to predict reaction times from model outputs will be referred to as 'Linear Fits' to avoid confusion with Language Models.

2.1 Syntactic Test Suites

We use sixteen test suites for syntactic generalization, adapted from Hu et al. (2020). Test suites consist of 20-25 items. Each item appears in four conditions, two grammatical and two ungrammatical.[2] Table 1 gives the name of each test suite, an example, as well as a tag, which we will use to refer to that suite in figures. When test suites have modifiers they always included distractors of the opposite grammatical category. For example singular reflexive anaphora sentences with subject relative clause modifiers would have a plural noun in the relative clause (e.g. *The bishop who likes the kings saw *themselves/himself in the mirror.*)

Following the logic from Hu et al. (2020), each test suite comes with two or more criteria, which specifies an inequality that should hold in a particular *critical region* if model behavior follows the rules of the relevant grammatical construction. Accuracy scores for each test suite are generated by computing the proportion of the time the inequality holds within the critical region, across items in a test suite. In Hu et al., test suites include criteria that correspond to 2-way contrasts between grammatical/ungrammatical conditions as well as 2x2 interactions between four conditions. We only look at the 2-way contrasts, here.

The incremental processing measure we derive from a language model to determine its accuracy according to a suite's inequality predictions is *surprisal*. Surprisal is the inverse log probability of a word given its context: $S(x_i) = -\log_2 p(x_i|x_1...x_{i-1})$, measured in bits. In this paper, we novelly extend the usage of these inequalities to determine a human *consistency score* for each test suite, by checking the mean reaction times for the various conditions of each item in the suite against the suite's criteria. For naturalistic corpus materials, the effect of surprisal on human reaction times has been shown to be linear (Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2020), motivating this usage of syntactic generalization criteria on human reading patterns. We use the same criteria as described in Appendix B of Hu et al. (2020).

To walk through a single test suite in detail, (1) gives an example of all four conditions of the *Main Verb / Reduced Relative Clause* suite, with critical regions underlined.

(1) a. The artist drawn a portrait was impressed with the work. [UN-REDUCED, UNAMBIGUOUS]
b. The artist that was drawn a portrait was impressed with the work. [REDUCED, UNAMBIGUOUS]
c. The artist painted a portrait was impressed with the work. [UN-REDUCED, UNAMBIGUOUS]
d. The artist that was painted a portrait was impressed with the work. [REDUCED, AMBIGUOUS]

The logic of the test suite relies on the fact that strings like *painted* are ambiguous between active past-tense main verbs and passive participles that introduce a reduced relative clause. On the other hand, verbs like *drawn* unambiguously introduce a reduced relative clause. If subjects believe that the ambiguous form of the verb introduces a main verb, they should find the critical-region verb *was impressed* surprising. That is, relative to the [REDUCED, AMBIGUOUS] conditions, not reducing the verb or not using an ambiguous verb should make the critical region less surprising (1 and 2 below). Furthermore, the effect of not reducing the relative clause should be smaller for unambiguous verbs than for ambiguous ones (3).

If we denote for convenience $S_x(w_i)$ as the surprisal of word w_i in the context of version x of a test suite item, then the following list outlines these three predictions as inequalities, which we used to determine accuracy scores on our test suites.

1. S_d(was impressed) < S_c(was impressed)
2. S_d(was impressed) < S_b(was impressed)
3. (S_d(was impressed) - S_c(was impressed)) < (S_b(was impressed) - S$_a$(was impressed))

To foreshadow our results, the **MVRR** panels of Figures 3 and in Appendix A show that all three of these criteria are met for most items both by all models and by human average reaction times. Unlike our other test suites, these predictions do not correspond to contrasts between sentences that vary based on their grammaticality, but rather on predictive processing that prefers the main-verb analysis for locally ambiguous strings.

2.2 The Interpolated Maze Task

Human reaction time data was collected via a novel implementation of the Maze Task (Forster et al., 2009) which we call the *Interpolated Maze*. In a maze task participants read through a sentence; at each index they are presented with two possible continuations, one word is a plausible next-word

[2] For the MVRR test suites, the 'ungrammatical' conditions are plausibly licensed by the grammar, but are unlikely. Following convention in linguistics, ungrammatical sentences will be marked with a *.

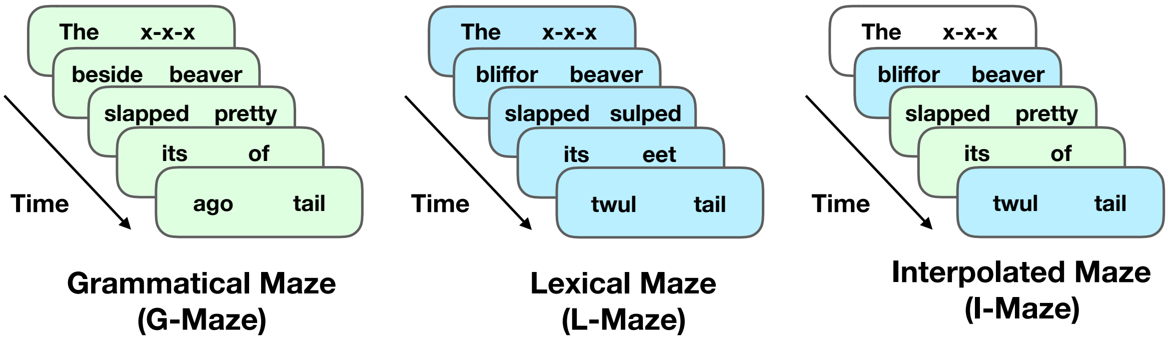

Figure 1: The Maze Task: Participants read the sentence word-by-word. At each index they must select the right continuation. For this study, we introduce the *Interpolated Maze*, which is a blend of G-Maze and L-Maze.

in the sentence and the other word is a distractor. Participants must select the correct continuation by pressing a key on their keyboard. Figure 1 shows a cartoon of this process for three variants of the Maze Task. In the G(rammatical)-Maze version, the distractor word is a word of English, only it does not constitute a grammatical continuation. In the L(exical)-Maze variant, the word is a non-English nonce word. If participants select the wrong continuation, the trial ends and they begin reading the next sentence. The time it takes participants to select the correct word by pressing a key has been shown to be a robust measure of incremental processing difficulty, with slowdowns occurring on target words instead of in subsequent spillover regions as is the case with other online processing measures such as self-paced reading (Boyce et al., 2020).

Of these two variants, G-Maze has been shown to produce higher sensitivity results than L-Maze (Boyce et al., 2020), however because each index must present one possible continuation, it cannot be used be used for items that have ungrammatical conditions. At the critical choice point, both the distractor and the continuation would be ungrammatical and participants would not know which continuation to select. To solve this problem we deploy a novel variant of the maze task called *Interpolated Maze*, or I-Maze. In I-Maze, we interweave G-Maze and L-Maze choices, with L-Maze distractors in critical regions where one of the conditions is ungrammatical. Participants are instructed to choose English words over nonce-words, thus making the 'right' choice in these regions unambiguous. In order not to clump L-Maze distractors only in critical regions, we randomly sample ∼25% of all other words and render them as L-Maze choices. For a full comparison of I-Maze, G-Maze and L-Maze see Vani et al. (2021). G-Maze distractors were generated with the scripts provided in Boyce et al. (2020), which uses a neural-network based language model to automatically generate high surprisal distractor words. Nonce words were generated with Wuggy (Keuleers and Brysbaert, 2010). Experiments were hosted on Ibex Farm (Drummond, 2013), with participants recruited on Amazon M-Turk. reaction time data for each item was collected from thirty separate participants.

2.3 Models Tested

JRNN is the 'BIG LSTM+CNN Inputs' from Jozefowicz et al. (2016). It was trained on the One Billion Word Benchmark (Chelba et al., 2013) with two hidden layers of 8196 units each and CNN character embeddings as input.

GRNN is the best-performing model described in the supplementary materials of Gulordava et al. (2018). It was trained on 90 million tokens of English Wikipedia with two hidden layers of 650 hidden units.

GPT-2 is the model presented in Radford et al. (2019), and was trained on 40GB of internet text. We use the version of GPT-2 available through the `Language Modeling Zoo` distribution[3]

RNNG (Dyer et al., 2016) jointly models a sentence as well as its syntactic parse. The model explicitly represents parse trees and composes partially built phrase structures. Models are supervised with Penn-Treebank style parses during training. We use the average of the three RNNG-BLLIP-LG models from Hu et al. (2020).

[3] `https://cpllab.github.io/lm-zoo/index.html#welcome-to-lm-zoo`

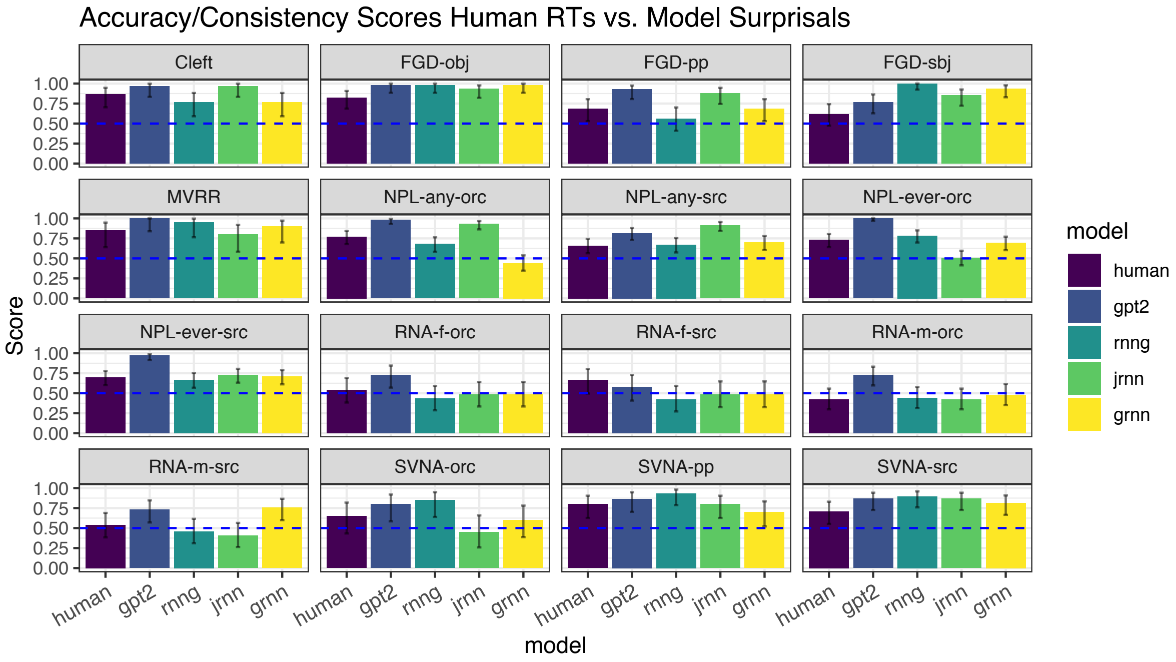

Figure 2: Comparison between human consistency scores and model accuracy scores. Averages are taken across all predictions within a test suite, error bars are 95% binomial confidence intervals. Scores are similar between humans and models

2.4 Addressing Two Possible Confounds

Before we turn to our results, we will briefly address two possible confounds with our methods: First, while it may be the case that the relationship between surprisal and reaction time is linear in most sentence areas, this linearity may break down in high surprisal regions regardless of the underlying grammaticality of the sentence. Thus, any potential badness of our linear fits in critical regions is an epiphenomenon of the fact that they were trained in regions where the linearity holds and tested in regions where it does not. While there is some evidence that the linear relationship between surprisal may flatten off in high surprisal regions for self-paced reading (see, e.g. Figure 1 in Wilcox et al. (2020)), data collected for Maze task for both GRNN and a large Transformer model shows that the linear relationship holds even in very high surprisal regions, even exceeding 20 bits (Boyce and Levy, 2020) (see, especially Figure 3).

The second confound has to do with the Interpolated Maze task. It may be the case that switching between tasks incurs a cognitive load, thus ungrammatical sentence regions might be read more slowly, but only because they are always associated with a switch from grammatical to lexical distractors. This could be worrisome, however we find that reaction times in non-critical regions for L-Maze decisions are actually slightly *faster* than G-Maze decisions ($p < 0.001$ by a t-test). Furthermore, all of our reported contrasts are between L-Maze items, so this is controlled for in our analyses.

3 Results

3.1 Test Suite Accuracy

In this section we discuss test suite accuracy scores, which are computed using the predictions associated with each test suite. For models, success on a prediction means that the model found material in a specified *critical region* more probable in the grammatical condition than the ungrammatical condition. For humans, a corresponding metric, *consistency scores*, report the proportion of times the critical region material was read more quickly in the grammatical condition than in the ungrammatical condition. Scores are calculated across the total number of items in a test suite. Because multiple subjects provided reaction time data for each item, we first average item-level data across all participants before calculating consistency scores.

The accuracy/consistency scores for each of our test suites can be seen in Figure 2. In this figure each facet represents the results from a single test suite, which aggregates across two or more predic-

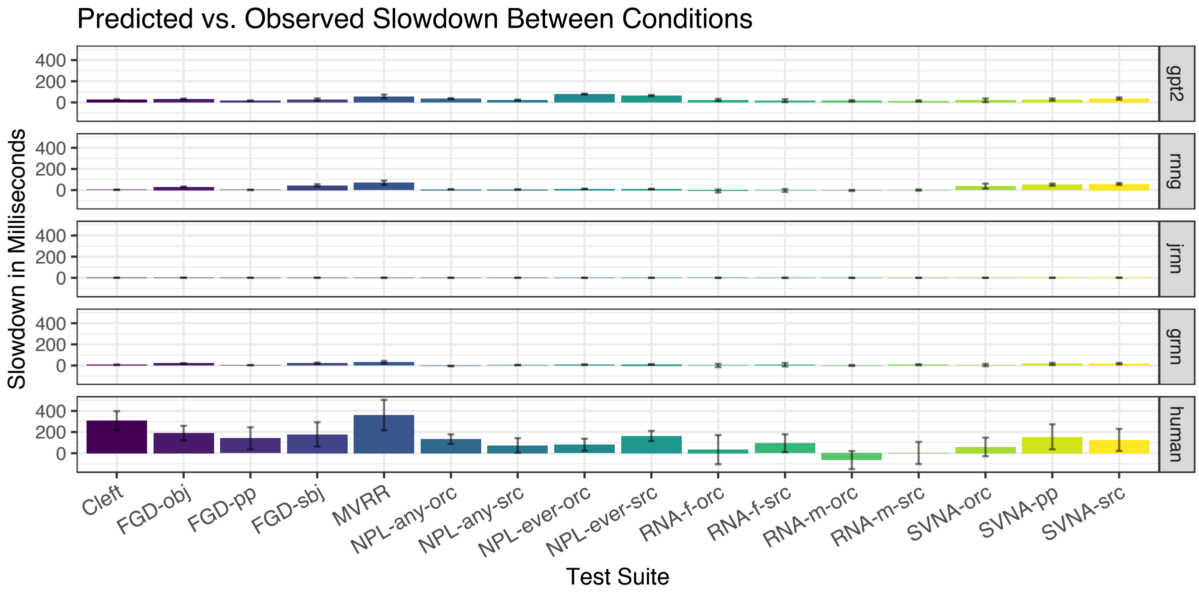

Figure 3: Comparison between human and (predicted) model reaction-time slowdows between grammatical and ungrammatical conditions. Averages are taken across all predictions within a test suite, error bars are 95% confidence intervals. Models systematically under-predict the observed slowdown.

tions. A full breakdown of test suite by prediction can be seen in Appendix B. Chance, which is 50% accuracy, is marked with a dashed blue line.

Humans perform above chance on 13/16 test suites. Human RTs are at or below chance for 3/4 of the Reflexive Anaphora agreement tests and the Subject-Verb Number Agreement with an Object Relative Clause modifier. For the Reflexive Anaphora tests, the low scores are driven by poor performance when the noun that must be matched is singular, such as in *The lawyer who the judges fear hurt herself/*themselves*. Notably, human reaction times for negative polarity items and for number agreement on verbs and reflexive pronouns are known to be susceptible to facilitatory interference effects from intervening attractors of the sort that are used in our test suites (Vasishth et al., 2008; Jäger et al., 2020). In general, human consistency scores in this study are below that reported in Marvin and Linzen (2018), who use an offline forced-choice paradigm, in which participants must judge which of two sentences sounds more natural. Nevertheless, for the vast majority of test suites, humans show robust sensitivity to the grammatical effects being tested, and failure is due to specific biases, such as the singular reflexive behavior discussed above, not general insensitivity to the manipulations.

Table 2 shows the cross-suite correlations between human consistency scores and model accuracy scores. The relatively strong correlation scores indicate that the strength of signal for a syntactic generalization in model surprisal differentials is predictive of the signal-to-noise ratio for the generalization in human reaction times.

Model	Correlation	p-value
GRNN	0.45	0.07
JRNN	0.68	< 0.01
GPT2	0.71	< 0.01
RNNG	0.65	< 0.01

Table 2: Correlations between model accuracy scores and human consistency scores across test suites.

3.2 Slowdown Between Conditions

In this section we turn to the size of the contrast between grammatical and ungrammatical conditions. For humans, this contrast indicates a slowdown, where critical regions of ungrammatical sentences are read more difficultly than their corresponding grammatical variants. For LMs, this contrast indicates a surprisal difference, where ungrammatical conditions are more surprising than their grammatical counterparts. Do differences in surprisal accurately predict the slowdowns observed in human reaction time data?

To derive a predicted reaction-time slowdown from the model surprisals, we followed the method-

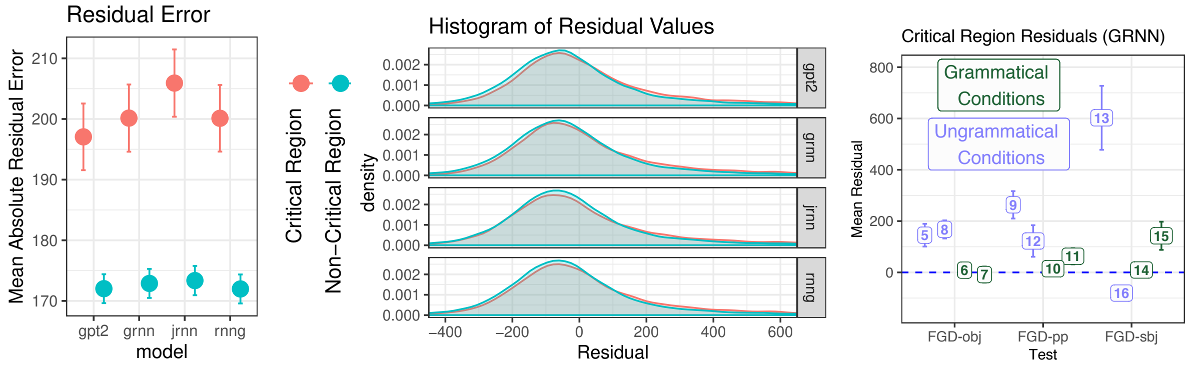

Figure 4: Residuals for reaction times in critical regions from a linear fit trained to predict reaction times from surprisal values in non-critical sentence regions. The left facet shows mean absolute residual error and the center shows a histogram of the raw values, with larger residuals in critical regions. The right facet shows a breakdown by condition for the Filler–Gap Dependency tests (GRNN model), with larger residual values in the ungrammatical conditions. For this plot, labels indicate condition name, with a reference provided in Appendix A. Error bars are 95% confidence intervals.

Model	Surprisal Estimate	p-value
GRNN	$8.8ms/bit$	< 0.001
JRNN	$0.5ms/bit$	< 0.05
GPT2	$12.0ms/bit$	< 0.001
RNNG	$19.0ms/bit$	< 0.001

Table 3: Surprisal Estimates from Linear Fits

ology outlined in Van Schijndel and Linzen (2018). This approach draws on the fact that the relationship between surprisal and human reaction time is linear across multiple orders of magnitude (Smith and Levy, 2013; Wilcox et al., 2020), including for Maze data (Boyce and Levy, 2020). For each LM, we trained a linear fit that predicts reaction time from surprisal value at the word-level. The model is fit on RTs from all L-Maze distractor trials, critical and non-critical region alike, and includes word frequency and word length as additional predictors, with random slopes for each item and each participant. The linear model's surprisal estimate, therefore, is the slowdown in processing time predicted for each bit of surprisal. We treat this number as a scalar and multiply it by the difference in surprisal between conditions to derive the total predicted slowdown due to syntactic violation from the language models. For all of our fits, we found a significant effect for all of our predictors. The estimates for each model's surprisal term are given in Table 3.

The results from this analysis can be seen in Figure 3, with the various test suites on the x-axis and observed or predicted slowdowns on the y-axis. As with accuracy scores, we average across predictions within each test suite. Humans demonstrate positive slowdowns in 11/16 test suites, with reflexive anaphora again proving the exception to the general trend. As is evident from the height of the bars, models systematically under-predict the slowdown observed in the human data. Models' predictions are outside of the 95% confidence intervals for the humans slowdowns in 7/16 test suites for GPT2, 8/16 for RNNG, 9/16 for GRNN and 12/16 for JRNN. The mean predicted difference between models and humans across all test suites is $95ms$ (GPT2), $107ms$ (RNNG), $117ms$ (GRNN) and $126ms$ (JRNN). These data indicate that models are less sensitive to the contrast between grammatical and ungrammatical conditions than are humans, at least in this controlled testing environment.

3.3 Residuals

In this section, we discuss a follow-up analysis conducted to validate the conclusion that models are under-predicting reaction times in critical regions. To do this, we train linear fits on data from the non-critical regions, and get their residuals on data from these regions as well the critical regions. The linear fits are exactly the same as the ones described in the previous section, except instead of being trained on both critical and non-critical L-Maze trials, they are trained on non-critical L-Maze trials alone. If the conclusion from the last section is correct, then we should see larger residuals for the critical-region data then for the non-critical region data.

The results from this analysis can be seen in the right and center facets of Figure 4. The left facet shows the mean absolute value of the residuals for each of our LMs, both for the critical and non-critical region. The center facet shows a histogram of the same data. From both plots it is clear that the critical region residuals are greater than the residuals computed for words in other regions of the sentence. From the histograms, we can see that the critical region residuals are systematically higher on average than the non-critical region residuals. This indicates that the models under-predict the RT values in the critical regions.

The difference between residuals provides additional evidence that models under-predict reaction times in critical regions compared to words in other parts of the sentence. However, it does not show that models under predict reaction times specifically for *ungrammatical* sentences. To investigate this, we break down average residual by condition, within each of our sixteen test suites. The full results for this breakdown can be seen in Appendix B, with the results for the Filler–Gap dependency tests for the GRNN model in the right facet of Figure 4.[4] Across all tests, we find that ungrammatical conditions show much higher residual error. The mean absolute value of the residual error is $163ms$ in grammatical conditions, but in ungrammatical conditions it is $244ms$. The values of the two conditions are significantly different ($p < 0.001$ by a t-test). Generally, residuals are largest for Cleft, Filler–Gap Dependency and MVRR suites, and smaller for suites that involve NPI Licensing, Anaphora agreement and Subject-Verb Number agreement. Human reaction-times are known to be susceptible to interference effects from distractors for these syntactic phenomena (Jäger et al., 2020), which may explain why residuals are smaller for these suites. Taken together this analysis demonstrates that model surprisal values specifically under predict human reaction times in ungrammatical critical regions, suggesting that they are less sensitive to syntactic violations than are humans.

4 Discussion

Our experiments have tackled the question of whether syntactic difficulty can be reduced to by-word probabilities by providing a comparison of Language Model and human behavior that is both incremental and targeted. Our methods build on those presented in Van Schijndel and Linzen (2018) and van Schijndel and Linzen (2020), but differ from theirs in a number of key respects, which we review briefly below to highlight to novel aspects of our own investigation. First, all of our test suites target grammatical/ungrammatical contrasts (except for the MVRR gardenpath test), whereas van Schijndel and Linzen test locally ambiguous sentence regions that (may) require re-analysis for proper processing. Second, we assess a broad range of grammatical violations across sixteen test suites that target seven distinct structures. Third, we deploy a novel measurement of processing time (*Interpolated Maze*), instead of self-paced reading. We fit our own linear models from the I-Maze data, and use a ms/bit scalar term derived from lexical distractor items. Finally, we provide a novel analysis that compares the residuals of linear fits between critical and non-critical regions, and we break down these residuals based on the grammaticality of the condition.

4.1 Model Comparison

While none of our models is able to capture humanlike sensitivity in ungrammatical critical regions, we do see some variation between them, with RNNG and GPT-2 in particular showing the most humanlike results. To compare model performance for accuracy scores (i.e. the results presented in Section 3.1), we fit pairwise logistic regression models, with the model class as the sole predictor, and random slopes for nested item/test suite combinations and predictions (this because predictions are shared across test suites of the same type). We find that GPT-2 performs significantly beter than both JRNN and GRNN ($p < 0.01$) and the contrast between RNNG and GRNN approaches significance ($p = 0.07$) None of the other pairwise comparisons are significant.

To compare model performance at predicting human slowdown in critical regions, we look at the difference in residual errors between the models from Section 3.3 in the critical regions. We fit liner regression models with the residual as predictor variable, nested item/test suite combinations, and *condition* as random slopes. We find a significant contrast between GPT-2 and JRNN ($p < 0.05$), with GPT-2 performing better, and a near-significant contrast between RNNG and JRNN ($p = 0.053$). Overall, these results support the conclusion that GPT-2 and RNNG have

[4] With the MVRR test suite, no conditions are technically ungrammatical, however we treat the *reduced_ambiguous* condition as ungrammatical for the purposes of this analysis.

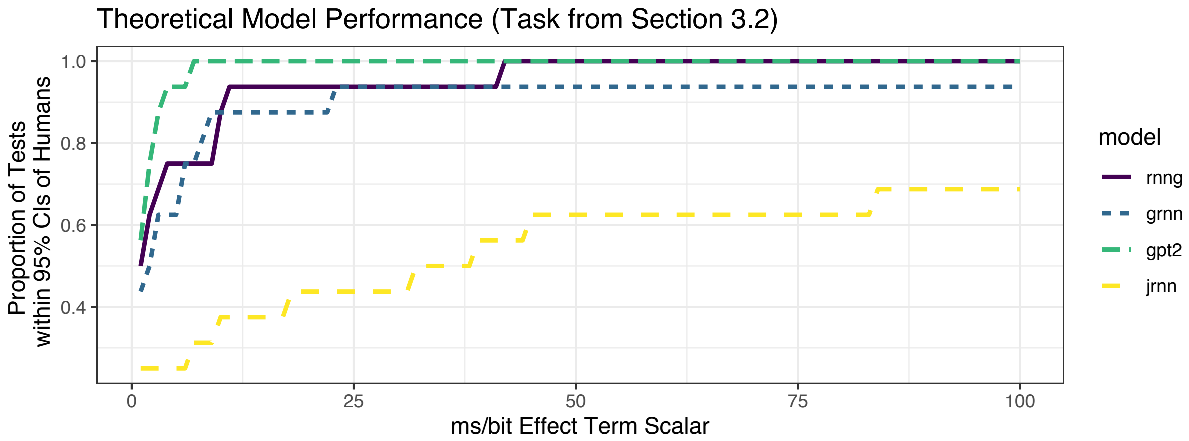

Figure 5: The effect of an additional ms/bit scalar term on model performance from tests in Section 3.2. Results indicate that both the RNNG and GRNN models could reach near human-like performance (within the human confidence intervals 90% of the time) when the scalar term is around 10.

a mild advantage over the other models. This is especially interesting for the RNNG model, given that it was trained on orders of magnitude less data than GPT-2.

4.2 Single Stage Models

For the last decade, a "single-stage" theory of incremental processing (Levy, 2008), in which word surprisal in a left-to-right language model (with a large or unlimited beam for models that explicitly represent multiple incremental parses) is the sole determinant of the processing difficulty that arises due to the relationship between a word and the context it appears in, has been a prominent candidate theory for both experimental (Staub, 2011) and computational (Frank and Bod, 2011) psycholinguistic investigations. Although such a "single-stage" can capture the *qualitative* difficulty patterns induced by garden-pathing and other grammar-based expectation violations (Hale, 2001; Levy, 2013), we now see that it *quantitatively* under-predicts the difficulty induced when grammatical expectation violations are involved, as measured by self-paced reading (van Schijndel and Linzen, 2020) and response times in the Maze task (here).

But just how bleak is the outlook for single-stage models? To investigate this, we re-analyze the results from Section 3.2 with theoretical model performance that includes an additional scalar term that corresponds with an increase in the slope for surprisal relative to that obtained from the fit to reaction times. The results in Figure 5. Here, the y-axis shows the proportion of tests for which the models are within the confidence intervals of human results, and the x-axis shows this scalar term. We find that models achieve 90% accuracy levels when the scalar term is 4 for GPT2, 11 for RNNG and 23 for GRNN. What this means is that if either the ms/bit scalar term, or the surprisal in ungrammatical conditions were (slightly under) an order of magnitude greater, then the models' performance would match humans.

While we agree with the assessment from van Schijndel and Linzen (2020) that these results pose a challenge for contemporary implemented models, we do not necessarily believe that they cannot be overcome within the framework of single-stage models, especially ones that are mediated by symbolic representations like the RNNG. Multiple options exist that could magnify surprisal values in locally ambiguous or ungrammatical regions, such as a reduced beam size (Roark, 2001) or particle filters (Levy et al., 2009). Taken together, these recent results highlight a key question for future research—what additional modeling mechanisms will be needed to accurately predict not only qualitative but also quantitative patterns of human difficulty in language processing.

Acknowledgements

RPL gratefully acknowledges NSF grant BCS-1551866, a Google Faculty Research Award, and funding from the MIT–IBM AI Lab.

Ethical Considerations

Data were collected under an Institutional Review Board (IRB) approved protocol for online human subject experimentation. Participants were compensated \$2.00 for their participation in I-Maze experiments. Experiments took ∼15 minutes, which meant participants were being compensated ∼\$8.00/hour. We chose this rate because it is slightly above federal minimum wage, which we take to be a fair baseline for compensation. All information associated with experimental participants was anonymized prior to analysis.

References

Veronica Boyce, Richard Futrell, and Roger P Levy. 2020. Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, 111:104082.

Veronica Boyce and Roger Levy. 2020. A-maze of natural stories: Texts are comprehensible using the maze task. *Proceedings of the Architectures and Mechanisms for Languages Processing Conference*.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Alex Drummond. 2013. Ibex farm. *Online server: http://spellout. net/ibexfarm*.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Kenneth I Forster, Christine Guerrera, and Lisa Elliot. 2009. The maze task: Measuring forced incremental sentence processing time. *Behavior research methods*, 41(1):163–171.

Stefan L. Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–834.

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

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*.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18.

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*.

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

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P Levy. 2020. A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692*.

Lena A Jäger, Daniela Mertzen, Julie A Van Dyke, and Shravan Vasishth. 2020. Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*, 111:104063.

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

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of lstms to understand negative polarity items. *arXiv preprint arXiv:1808.10627*.

Emmanuel Keuleers and Marc Brysbaert. 2010. Wuggy: A multilingual pseudoword generator. *Behavior research methods*, 42(3):627–633.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Roger Levy. 2013. Memory and surprisal in human sentence comprehension. In Roger P. G. van Gompel, editor, *Sentence Processing*, pages 78–114. Hove: Psychology Press.

Roger P Levy, Florencia Reali, and Thomas L Griffiths. 2009. Modeling the effects of memory on human online sentence processing with particle filters. In *Advances in neural information processing systems*, pages 937–944.

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.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

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.

Brian Roark. 2001. Probabilistic top-down parsing and language modeling. *Computational linguistics*, 27(2):249–276.

Marten van Schijndel and Tal Linzen. 2020. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Adrian Staub. 2011. Word recognition and syntactic attachment in reading: Evidence for a staged architecture. *Journal of Experimental Psychology: General*, 140(3):407.

Marten Van Schijndel and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In *CogSci*.

Pranali Vani, Ethan Gotlieb Wilcox, and Roger Levy. 2021. Using the interpolated maze task to assess incremental processing in english relative clauses. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Shravan Vasishth, Sven Brüssow, Richard L Lewis, and Heiner Drenhaus. 2008. Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, 32(4):685–712.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

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*.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

A Consistency/Accuracy Scores by Prediction

Figure 6 gives accuracy scores for humans and LM models, broken down by individual predictions. Predictions are taken from (Hu et al., 2020), outlined in their Appendix B. Prediction names correspond to the licensed element of the sentence, so *sing_match_prediction* for reflexive anaphora licensing corresponds to the contrast where *himself* or *herself* is grammatical (as opposed to *themselves*). Accuracy/consistency scores are similar between humans and models for cleft structures, filler–gap dependencies (except for *subject* tests, which we discuss below), MVRR gardenpath and Subject Verb Number Agreement suites. In the rest of this appendix, we focus in on structures that show different accuracy/consistency score patterns for humans and models.

For filler–gap dependency tests, the human data differs from the model data when there is a gap in the *subject* position (FGD-sbj test). In this case, both achieve relatively high scores for the *wh prediction* (yellow bars), but lower scores filled-gap prediction (`I know *who/that my mother...`). (It should be noted that this contrast is not one strictly of grammaticality in the critical region, as the sentence could be felicitously completed by a gap in the object position.) This behavior is in perfect alignment with the large amount of data demonstrating that English speakers take longer processing object gaps over subject gaps, and suggests that such expectations are weaker in our neural models.

Turning to NPI and anaphor licensing, we see a consistent pattern of difference between humans and models. For the NPI tests, models perform much worse than humans at the swap_intervener predictions (`No senator that the lawyer liked ... ever/any` vs. `The senator that no lawyer liked ... ever/any`), whereas human participants performed about as well on these tests as on the others. For reflexive anaphora licensing, human performance is worse for the *singular* predictions, regardless of the gender of the pronoun, indicating a plural bias across the board. For models, this is true only for the feminine pronoun (*herself*), and the difference in accuracy is much greater than the human difference in consistency scores. When the masculine version of the pronoun is used, models show similar

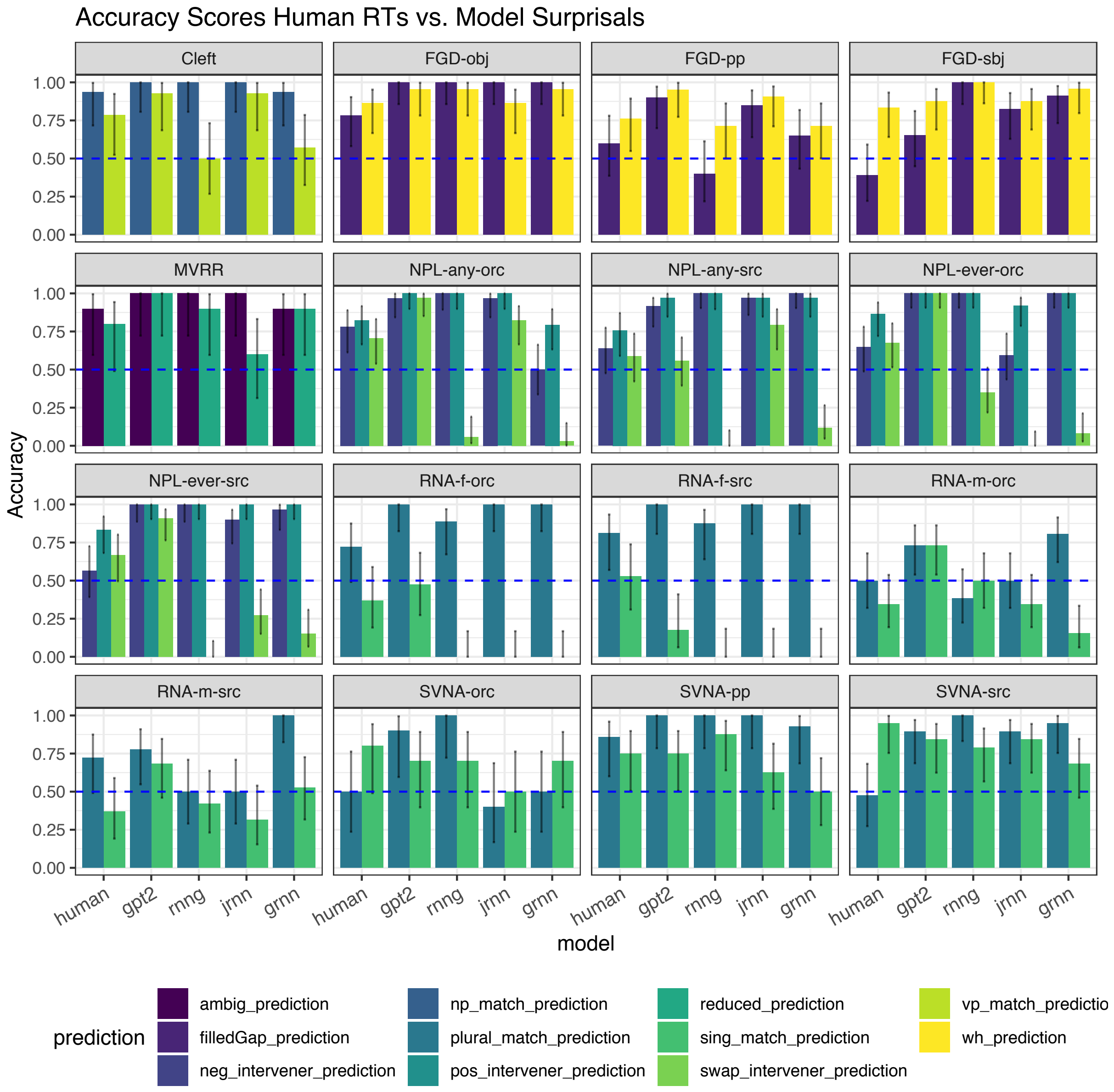

Figure 6: Test suite accuracy / consistency scores broken down by individual predictions.

scores for both the *singular* and *plural* predictions. This pattern is consistent with a plural bias in humans, but a bias against specifically the feminine (singular) form of the pronoun in models.

B Linear Fit Residuals by Condition

Table 4 gives a breakdown of all test suite conditions, with an example and a tag used for labeling for the left panel of Figure 4 in the main text and for the figures in this appendix. Ungrammatical conditions are marked with a star. Figure 7 shows the residuals from our linear fits for each condition/test suite pair. See the figure caption for more detail.

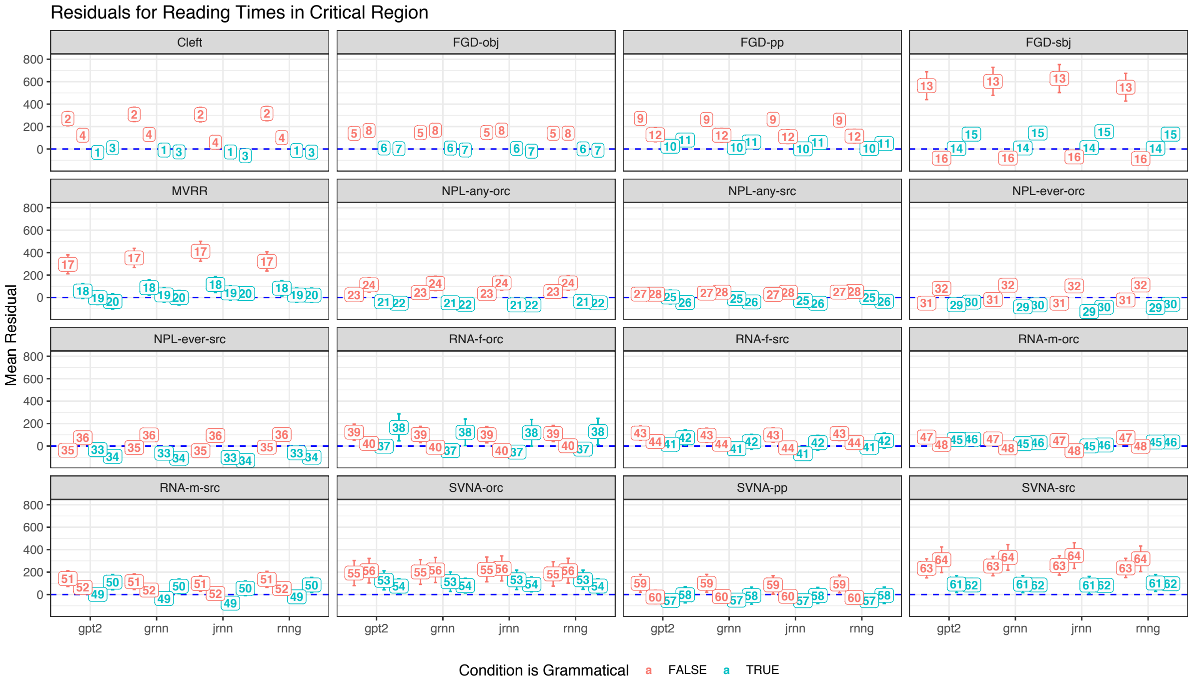

Figure 7: Residuals for predicted reaction times in critical regions, from a linear fit trained to predict reaction times from surprisal values in non-critical regions. Labels indicate condition name, with a reference provided in Appendix A. Error bars are 95% confidence intervals. Across the majority of test suites, ungrammatical conditions show larger residuals, indicating that they are predicted less well by LM surprisal values.

Condition Label	Test Suite Name	Condition Name	Example
1	Cleft	np-match	What she spied was the giraffe
2	Cleft	np-mismatch	*What she spied was see the giraffe
3	Cleft	vp-match	What she did was see the giraffe
4	Cleft	vp-mismatch	*What she did was see the giraffe
5	FGD-obj	that-gap	*I know that my mother sent — to Taylor yesterday.
6	FGD-obj	that-nogap	I know that my mother sent the present to Taylor yesterday.
7	FGD-obj	what-gap	I know what my mother sent — to Taylor yesterday.
8	FGD-obj	what-nogap	*I know what my mother sent the present to Taylor yesterday.
9	FGD-pp	that-gap	*I know that my mother sent the present to – yesterday.
10	FGD-pp	that-nogap	I know that my mother sent the present to Taylor yesterday.
11	FGD-pp	what-gap	I know who my mother sent the present to — yesterday.
12	FGD-pp	what-nogap	*I know who my mother sent the present to Taylor yesterday.
13	FGD-sbj	that-gap	*I know that — sent the present to Taylor yesterday.
14	FGD-sbj	that-nogap	I know that my mother sent the present to Taylor yesterday.
15	FGD-sbj	what-gap	I know who — sent the present to Taylor yesterday.
16	FGD-sbj	what-nogap	*I know who my mother sent the present to Taylor yesterday.
17	MVRR	reduced-ambig	The ship sunk the the storm carried treasure.
18	MVRR	reduced-unambig	The ship steered in the storm carried treasure.
19	MVRR	unreduced-ambig	The ship that was sunk in the storm carried treasure.
20	MVRR	unreduced-unambig	The ship that was steered in the storm carried treasure.
21	NPL-any-orc	neg-neg	No senator that no journalist likes has gotten any votes.
22	NPL-any-orc	neg-pos	No senator that the journalist likes has gotten any votes.
23	NPL-any-orc	pos-neg	*The senator that no journalist likes has gotten any votes.
24	NPL-any-orc	pos-pos	*The senator that the journalist likes has gotten any votes.
25	NPL-any-src	neg-neg	No senator that likes no journalists has gotten any votes.
26	NPL-any-src	neg-pos	No senator that likes the journalists has gotten any votes.
27	NPL-any-src	pos-neg	*The senator that likes no journalists has gotten any votes.
28	NPL-any-src	pos-pos	*The senator that likes the journalist has gotten any votes.
29	NPL-ever-orc	neg-neg	No senator that no journalist likes has ever won.
30	NPL-ever-orc	neg-pos	No senator that the journalist likes has ever won.
31	NPL-ever-orc	pos-neg	*The senator that no journalist likes has ever won.
32	NPL-ever-orc	pos-pos	*The senator that the journalist likes has ever won.
33	NPL-ever-src	neg-neg	No senator that likes no journalists has ever won.
34	NPL-ever-src	neg-pos	No senator that likes the journalists has ever won.
35	NPL-ever-src	pos-neg	*The senator that likes no journalists has ever won.
36	NPL-ever-src	pos-pos	*The senator that likes the journalist has ever won.
37	RNA-f-orc	match-plural	The queens who the dukes mistrust saw themselves in the mirror.
38	RNA-f-orc	match-sing	The queen who the duke mistrusts saw herself in the mirror.
39	RNA-f-orc	mismatch-plural	*The queens who the dukes mistrust saw herself in the mirror.
40	RNA-f-orc	mismatch-sing	*The queen who the dukes mistrust saw themselves in the mirror.
41	RNA-f-src	match-plural	The queens who hunted the rabbit saw themselves in the mirror.
42	RNA-f-src	match-sing	The queen who hunted the rabbits saw herself in the mirror.
43	RNA-f-src	mismatch-plural	*The queens who hunted the rabbit saw herself in the mirror.
44	RNA-f-src	mismatch-sing	*The queen who hunted the rabbits saw themselves in the mirror.
45	RNA-m-orc	match-plural	The dukes who the dukes mistrust saw themselves in the mirror.
46	RNA-m-orc	match-sing	The duke who the duke mistrusts saw himself in the mirror.
47	RNA-m-orc	mismatch-plural	*The dukes who the dukes mistrust saw himself in the mirror.
48	RNA-m-orc	mismatch-sing	*The duke who the dukes mistrust saw themselves in the mirror.
49	RNA-m-src	match-plural	The dukes who hunted the rabbit saw themselves in the mirror.
50	RNA-m-src	match-sing	The duke who hunted the rabbits saw himself in the mirror.
51	RNA-m-src	mismatch-plural	*The dukes who hunted the rabbit saw himself in the mirror.
52	RNA-m-src	mismatch-sing	*The duke who hunted the rabbits saw themselves in the mirror.
53	SVNA-orc	match-plural	The lawyers that helped the mayor are organized.
54	SVNA-orc	match-sing	The lawyer that helped the mayors is organized.
55	SVNA-orc	mismatch-plural	*The lawyers that helped the mayor is organized.
56	SVNA-orc	mismatch-sing	*The lawyer that helped the mayors are organized.
57	SVNA-pp	match-plural	The lawyers that the mayor helped are organized.
58	SVNA-pp	match-sing	The lawyer that the mayors helped is organized.
59	SVNA-pp	mismatch-plural	*The lawyers that the mayor helped is organized.
60	SVNA-pp	mismatch-sing	*The lawyer that the mayors helped are organized.
61	SVNA-src	match-plural	The lawyers next to the mayor are organized.
62	SVNA-src	match-sing	The lawyer next to the mayors is organized.
63	SVNA-src	mismatch-plural	*The lawyers next to the mayor is organized.
64	SVNA-src	mismatch-sing	*The lawyer next to the mayors is organized.

Table 4: Conditions for each of the test suites assessed in this paper, with a tag (used for labeling in Figure 7) and an example. Ungrammatical sentences are marked with a star (∗)