Of Models and Men: Probing Neural Networks for Agreement Attraction with Psycholinguistic Data

Maxim Bazhukov🤖**, Ekaterina Voloshina, Sergey Pletenev**✍️👨‍🎓🤖
Arseny Anisimov🤖 **Oleg Serikov**🖥️**, Svetlana Toldova**🤖
🤖HSE University, 👨‍🎓AIRI, ✍️Skoltech, 🖥️KAUST

Abstract

Interpretability studies have played an important role in the field of NLP. They focus on the problems of how models encode information or, for instance, whether linguistic capabilities allow them to prefer grammatical sentences to ungrammatical. Recently, several studies examined whether the models demonstrate patterns similar to humans and whether they are sensitive to the phenomena of interference like humans' grammaticality judgements, including the phenomenon of agreement attraction.

In this paper, we probe BERT and GPT models on the syntactic phenomenon of agreement attraction in Russian using the psycholinguistic data with syncretism. Working on the language with syncretism between some plural and singular forms allows us to differentiate between the effects of the surface form and of the underlying grammatical feature. Thus we can further investigate models' sensitivity to this phenomenon and examine if the patterns of their behaviour are similar to human patterns. Moreover, we suggest a new way of comparing models' and humans' responses via statistical testing. We show that there are some similarities between models' and humans' results, while GPT is somewhat more aligned with human responses than BERT. Finally, preliminary results suggest that surface form syncretism influences attraction, perhaps more so than grammatical form syncretism. [1]

1 Introduction[2]

With the fast development of large language models (LLMs), interpretability has become (Belinkov et al., 2023) an important issue in Natural Language Processing. Interpretability studies aim to explain what LLMs learn during pre-training, for instance, whether they pick up factual information or develop language skills. One of the promising directions of interpretability research is comparing model responses to human acceptability judgements (Lau et al., 2017; Warstadt et al., 2018). A case in point is the research on the phenomenon of agreement attraction.

In this paper, we investigate models' sensitivity to agreement attraction in light of morphological syncretism. We design a probing experiment based on the data from a previous psycholinguistic experiment on humans (Slioussar, 2018). The focus is on Russian, a language with rich morphology, which allows us to investigate agreement attraction interacting with different types of *syncretism*. *Syncretism* is a surface formal identity of grammatically distinct forms like genitive singular *polja* 'of field (GEN.SG)' and nominative plural *polja* 'fields (NOM.PL)' being formally identical. This kind of identity occurs in genitive case and it differs from accusative case syncretism of some other nouns, where it is simply the nominative and accusative case forms that coincide (separately in singular and in plural): *lug* 'meadow(NOM.SG=ACC.SG)' and *luga* 'meadows(NOM.PL=ACC.PL)'. Distinguishing accusative and genitive syncretisms is a unique setup that helps to disentangle the effects of structure (underlying features) from the effects of surface forms.

(1) a. *Trass-**a*** *čerez* *polje* *byl-**a*** *nov-**oj***
path-SG across field be.PST-SG new-SG
'The highway across the field was new'

b. *Trass-**y*** *čerez* *polje* *byl-**i*** *nov-**ymi***
path-PL across field be.PST-SG new-PL
'The highways across the field were new'

[1] The code for the experiments is available here: https://github.com/bamaxi/agreement-probing.

[2] For brevity we use the following abbreviations (glosses): SG – singular number; PL – plural number; NOM – nominative case; ACC – accusative case; GEN – genitive case; genders: M – masculine, F – feminine, N – neuter

Proceedings of the 28th Conference on Computational Natural Language Learning, pages 280–290
November 15-16, 2024 ©2024 Association for Computational Linguistics

Agreement is a grammar rule where grammatical features like number or gender of one linguistic element, the controller, license corresponding features on a syntactically related element, the target. In (1a) singular subject *trassa* is the controller and requires singular on the verb 'be' and adjective. Similarly, in (1b), plural *trassy* requires plural. This type of phenomena is well acquired by people (Guasti, 2017), who easily recognize errors in agreement (see e.g. (Slioussar, 2018)). However, the task becomes more complicated if the controller and the target of agreement are separated by some linguistic material. It is yet more complicated, if the surface form of the intervening material coincides with the form of a potential controller due to *syncretism*. The agreement errors are not recognised so easily in this case. The higher acceptability of incorrect sentences of this type is called *agreement attraction* and has been under research in psycholinguistics (Wagers et al., 2009) and in NLP. In NLP similar studies have already been widely conducted on the English language, revealing the inner workings of language models (Gulordava et al., 2018; Arehalli and Linzen, 2020), as well as the relationship between model errors and human errors (Linzen and Leonard, 2018).

We tackle a more complex question of whether and how the models parse two kinds of syncretism that could potentially cause agreement attraction for Russian. This allows us to finely distinguish the effect of the surface form from the effect of the underlying syntactical structure. We employ the data from a psycholinguistic study of (Slioussar, 2018), measuring Russian speakers reading times and cloze test completion. This further allows us to compare model's behaviour to that of humans.

Our contributions can be stated as following:

- We probe models on the task of agreement attraction with a new type of data. While recent studies were done for English, we work with a morphologically rich language with case-number syncretism, namely Russian;
- We compare the effect of syncretism to the effect of the underlying grammatical features
- We supply linguistic research with extra-human knowledge, showing to what extent neural networks' linguistic capabilities are similar to those of humans on the example of agreement attraction phenomenon;
- We propose a new way of comparing models' responses to the results of psycholinguistic experiments, as we perform more robust statistical analysis.

2 Related Work

2.1 Probing methodology

The interpretation of behaviour and learned representations of language models has been studied extensively. Belinkov et al. (2020) suggests classifying probing methods into *structural* and *behavioural*. *Structural* methods involve a diagnostic classifier, i.e. a simpler model, such as logistic regression, trained atop of embeddings from a bigger model. Such methods were criticised (Hewitt and Liang, 2019; Voita and Titov, 2020) for overrelying on an external classifier: it is not clear if the overall results of the studies depend on how well a model encoded linguistic information, and not on how well a classifier has been trained. *Behavioural* methods, on the other hand, involve no such external classifier and exploit models' inherit architecture. For example, Salazar et al. (2020) adapt masked language modelling task to probe internal linguistic knowledge of BERT.

2.2 Acceptability judgements

In linguistic theorizing, human acceptability judgements are an important tool. These are scores proxying grammaticality of the sentences (Chomsky, 1965; Schütze, 1996), binary (acceptable / unacceptable) or scalar. These were picked up in NLP (Lau et al., 2017) and, among other things, led to the creation of acceptability datasets like (Warstadt et al., 2018). Similarly, Warstadt et al. (2020) introduce a probing suite based on minimal pairs of grammatical and ungrammatical sentences. The suite covers several semantic, morphological and syntactic phenomena, such as negative polarity items, agreement and verb conjugation. It is shown that various behavioural model metrics can be chosen as analogues to human acceptability scores to establish preference of one sentence over another, and Warstadt et al. (2020) choose to compare full sentence likelihood. A similar work was recently done for Russian by Taktasheva et al. (2024). Indeed, this benchmark included sentences with attractor under subject-predicate agreement phenomenon, and models scored lower on such sentences than on similar sentences with no attractor. Our present work differs in that we compare human and models' performance on *psycholinguistic data*, designed for controlled experimental studies on human, with focus on syncretism-grammar comparison.

2.3 Psycholinguistics and neural networks

Since interpretation has become an important part of NLP research, several works have adapted psycholinguistic data to study how models acquire language. For example, Li et al. (2021) use psycholinguistic stimuli to study the effect of surprisal in RoBERTa layerwise showing that the best performing model shows surprisal already in the early layers. Other works adapt psycholinguistic concepts for better explanation of language model behaviour. Sinclair et al. (2022) use the effect of priming, studied earlier for humans, to see what can affect LLM's responses.

Other works directly or indirectly compare results of the models to human responses. Ettinger (2020) introduces a suite of several tasks taken from psycholinguistics to evaluate linguistic abilities of BERT. The author compares humans and the model on the basis of surface responses, such as sentence completion. Similarly, Li et al. (2022) adapt experiments based on the theory of Construction Grammar to study how different constructions are perceived by humans and models showing that transformers can detect constructions. They compare how the results of humans differ from the results of neural networks on such tasks as sorting preferable constructions. Wilcox et al. (2021) compare models' responses to human reaction time for a suite of syntactic tasks. They show that models resemble humans in their predictions although they do not achieve human-like level. Lampinen (2022) provides detailed discussion of how using proper psycholinguistic analysis of human evaluation allows drawing clearer insights from comparing LLMs to humans while bringing up the question of fair comparison of human and model responses.

2.4 Studies of attraction in agreement

Agreement is a phenomenon of licensing grammatical features like number or gender by one linguistic element, the controller, on another syntactically related element, the target. In general, while proper agreement requires understanding underlying hierarchic structure, subject-verb agreement is acquired early by human speakers (Guasti, 2017). Nonetheless, agreement is vulnerable to errors, particularly in the presence of "attractors" – subject noun dependents that are not subjects, but could be erroneously construed as subjects (see 2, 3 below).

(2) a. *The key to the cabinets were rusty
b. **The key to the cabinet were rusty

(3a) *_Trass-a čerez polj-a byl-i nov-ymi_
path-SG across field-PL be.PST-PL new-PL
'The highway across the fields were new'

(3b)**_Trassa čerez pol-e byli novymi_
path-SG across field-SG be.PST-PL new-PL
'The highway across the field were new'

Both sentences have longer reading times compared to fully grammatical sentences. However, sentences (2a) and (3a) show a reduced effect due to the presence of attractor nouns (*cabinets* and *polja* 'fields'). Here, these nouns could be construed as subjects and underlined parts could be proper sentences (see also Figure 1). This creates an illusion of grammaticality and mitigates the processing difficulty arising from the actual violation of grammar. Attraction of agreement is thus a grammatical notion, although similar interference effects are discussed for semantics, too (Timkey and Linzen, 2023). Hierarchy understanding by the models has been studied extensively for English (Gulordava et al., 2018; Arehalli and Linzen, 2020) and agreement attraction in particular has been compared in models and humans in a work similar to ours (Arehalli and Linzen, 2020).

One of the main sources of our data comes from Slioussar (2018). This study explores the role of *syncretism* (morphological ambiguity) in inducing attraction errors in number agreement, in Russian speakers. *Syncretism* is a phenomenon where two distinct moprhological categories are realized in the same way (Caha, 2019; Baerman et al., 2005). Unlike English, Russian nouns inflect for two categories: number and case, thus could potentially exhibit syncretism. Indeed, genitive singular *polja* 'of field (GEN.SG)' and accusative plural *polja* 'fields (ACC.PL)' are formally the same. Both are, in turn, identical to nominative plural *polja* 'fields (NOM.PL)' (all of these are, of course, distinguished in a context). Slioussar (2018) shows that surface syncretism in itself, independently of the underlying grammatical number feature, explains attraction effects. Thus ACC.PL and GEN.SG, surface forms both identical to what plural subject would be (=NOM.PL), show attraction effects, although GEN.SG is underlyingly singular (which is deducible from the syntactic structure of the full sentence). Crucially, Slioussar (2018) believes such data to be difficult for existing theories of attraction. We test whether the effect holds for models.

3 Experimental Setup

We study the attraction phenomenon in LLMs and compare it to human data available from (Slioussar, 2018). In these experiments, humans' reading time has been measured in relation to the grammatical pattern of the sentence. We follow this approach in our experimental setting, yet we also propose a model-specific interpretability analysis. We reproduce the reading time analysis performed on humans' data, by introducing the readability-like metrics for LLM. We offer deeper insights into how LLMs process sentences of every grammatical pattern, by analyzing their attention maps.

Our statistical analysis methodology mostly follows the one of Slioussar (2018), with a few changes made for the sake of results' interpretability. Since we test the models on the exact same data on which humans have been tested, we manage to avoid uneven comparisons, yet support the theoretical findings of the original work.

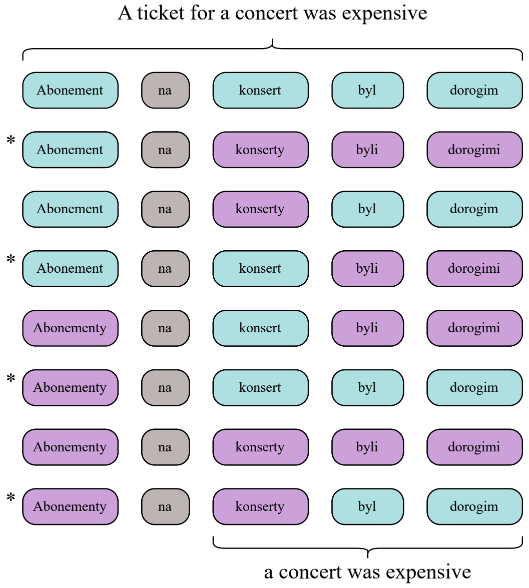
3.1 Models

We work with transformer-based models of different architectures: ruBERT[3], an encoder-only model, and ruGPT[4], a decoder-only architecture.

ruBERT (Kuratov and Arkhipov, 2019) was trained on the Russian part of Wikipedia and news data with pretraining objectives of Masked Language Modelling (MLM) and Next Sentence Prediction (NSP), following the original BERT architecture (Devlin et al., 2019).

ruGPT-3.5 (Zmitrovich et al., 2024) was trained on data from various domains (Wikipedia, books, and news) with a language modelling pretraining objective. The model is based on the original architecture of the GPT-3 model (Brown et al., 2020).

3.2 Data



Abonement na konsert byl dorogim
* Abonement na konserty byli dorogimi
Abonement na konserty byl dorogim
* Abonement na konsert byli dorogimi
Abonementy na konsert byli dorogimi
* Abonementy na konsert byl dorogim
Abonementy na konserty byli dorogimi
* Abonementy na konserty byl dorogim

Figure 1: Example of one set of data: Each sentence exists in 8 variants, formed by different combinations of the number of the subject (the first word), the attractor (third word), and the predicate (fourth and fifth words). Words in the singular are highlighted in blue, and words in the plural are highlighted in purple. Sentences marked with an asterisk ($*$) are ungrammatical.

To compare how perception of the attraction phenomenon differs in humans and models, we use the dataset prepared for a psycholinguistic study by Slioussar (2018), provided by the author upon our request.

The dataset includes in total 80 sentences with subject-verb agreement, full text is available for 64 of them. All our experiments with the models use this subset of 64 sentences, while fuller 80 sentence data is available for human response times (on these see the Section 4.3). All the sentences had the same syntactic structure: `subject + attractor + verb + other_verb_dependents`. The attractor is either in accusative case or in genitive case, splitting the data in halves. Also, in each of the 64 sentences, the subject, the attractor and the verb can have a singular or a plural marker amounting to 8 variants, as Figure 1 illustrates for a

[3] https://huggingface.co/DeepPavlov/rubert-base-cased

[4] https://huggingface.co/ai-forever/ruGPT-3.5-13B

condition with an attractor in accusative case. The total number of items is thus $64 * 8 = 512$. Notably, in all sentences the predicate is an adjective with *byt'* 'to be', the verb under examination, as an auxillary. We concur that using single lemma limits empirical coverage, but here it facilitates comparison.

As mentioned, sentences of Slioussar (2018) each belong to one of the two types. In half the sentences, the context demands that attractor be in the accusative case and in the other half that it be in the genitive case. Such setup allowed Slioussar to disentangle the effects of the underlying grammatical number from the effects of the surface form. This is because nominative plural is the form that could attract predicate agreement, and *accusative plural* is syncretic (has the same surface form) with nominative plural, while the reverse is true for genitive: it is *genitive singular*, that is syncretic with plural, while genitive plural is not (see Examples 4, 5). This is the unique property allowing to distinguish attraction by grammatical features and attraction by surface form. If attraction errors pattern the same way in accusative as in genitive, that would mean that only grammatical number is important. On the other hand, if these patterns were different depending on the case, surface form must matter too.

(4) ACCusative: ACC.**PL** = NOM.**PL**, ACC.SG ≠ NOM.PL

a. *tropinka cherez lug*[ACC.SG] *byla/*byli*
'a path trough the meadow was/*were'

b. *tropinka cherez* ***luga***[ACC.PL] *byla/*byli*
'a path trough the meadow was/*were'

c. ***luga***[NOM.PL] *byli*
'the meadows were'

(5) GENitive: GEN.PL ≠ NOM.PL, GEN.**SG** = NOM.**PL**

a. *korobka dlya* ***kraski***[GEN.SG] *byla/*byli*
'a box for the paints was/*were'

b. *korobka dlya krasok*[GEN.PL] *byla/*byli*
'a box for the paints was/*were'

c. ***kraski***[NOM.PL] *byli*
'paints were'

3.3 Methods

To evaluate models' behavior on the agreement attraction, we collect vectors representing model's activity when processing each of these 512 items. Then, inspired by (Slioussar, 2018), we employ statistical analysis to learn if observed features somehow reflect the sentence structure.

We hypothesize that eight groups of sentences can be meaningfully ranked by model's perplexity: sentences where attraction does happen as described above (e.g., 2a, 33a, 4b) should be more natural than the respective purely ungrammatical variants (2b, 33b, 4a), but less natural than correct sentences (6). Moreover, this effect should be reflected in human reading times.

(6) Predicted ranking of sentence types:
grammatical > attractor > ungrammatical

Most importantly, we expect to see one of three scenarios Slioussar (2018) outlined regarding the distinction between syncretism and underlying features. To test our hypotheses, we use two methods: perplexity-based and attention-based methods described below.

3.3.1 Estimation of models' certainty

Due to the differences in architectures and objectives of ruGPT and ruBERT, a direct comparison of the models' performance is not feasible. As the analysis of human behaviour in Slioussar (2018) was focused on word-level reading times, our analysis also focuses on word-level rather than sentence-level predictions.

In general, we want to estimate for each item how likely the verb is, given a prefix of subject and attractor (for grammatical items this is the correct verb form that agrees well with subject, and for ungrammatical ones — an incorrect form that does not). Such approach has already been shown to be effective for the study of attraction in GPT-like models (Arehalli and Linzen, 2020). Since this does not translate straightforwardly to BERT, to facilitate the comparison we establish the following methodological adjustments:

- **ruGPT**: we calculate the logarithmic probability of the first verb after the attractor word as an estimate of the model's generation.

$$Score_{GPT}(X) = log p_\theta(x_{i_{verb}}|x_{<i_{verb}})$$

where $x_{<i_{verb}}$ is tokens before verb.

- **ruBERT**: we use a masked language modeling approach. Specifically, we mask all tokens succeeding the attractor word. The generation estimate is then determined by subtracting the probability of the first masked token from the overall masked sequence probability.

$$Score_{BERT}(X) = logp_{\theta}(x_{verb}|context)$$

where $context$ is left part of the sentence $(x_0, x_1, ..., x_{verb-1})$.

This approach allows for a relative comparison of GPT and BERT's generation capabilities, focusing on the influence of the attractor word on subsequent word prediction, despite their distinct architectures and training objectives.

Both score estimates based on probability for BERT and GPT are naive implementations in this case. They may not work for other models or words (Kauf and Ivanova, 2023) (Pimentel and Meister, 2024). For some models word prediction estimation is made more difficult due to the tokenization step, where target word may be split into several tokens. In our case, however, we only predict one of four forms of 'to be' word: $byl(was)_{SG,masc}$ and all its variations $byla(was)_{SG,femn}$, $bylo(was)_{SG,neut}$, $byli(were)_{PL}$, which are tokenized as a single token for GPT and BERT.

3.3.2 Appoximating effect from a subject and an attractor

Apart from perplexity, we extract attention head projections and compare the attention distributions between different types of sentences. We take attention scores from each head and layer and then extract attention used to predict a predicate (an auxilary verb and an adjective) that comes from a subject and from an attractor. Therefore, we get two arrays representing attention from a predicate on a subject and an attractor respectively. These scores are averaged across attention heads for each layer. In other words, we calculate how much impact the subject had on prediction of a predicate and how much impact the attractor had on prediction of the same predicate. To compare the results on different sentence types, we use Student's T-test with Bonferroni correction.

4 Results and Discussion

4.1 How models perceive different types of ungrammatical sentences

To check whether models are sensitive to agreement errors in general, we evaluate their quality with adapted masked language model scoring (Salazar et al., 2020): we first calculate the scores (see Section 3.3.1) for each of our sentences and then we compare two sentences (grammatical and ungrammatical) that share the same subject and attractor and differ only in the number of the predicate. The sentence is grammatical if the number of the subject and the predicate match. We count the model as answering correctly, if the score for the grammatical sentence is higher than for the ungrammatical sentence. The results are summarised in Table 1, with the results for humans taken from the experiment 2 in (Slioussar, 2018) where participants were asked to complete a sentence. The tasks for models and humans are rather distinct and the data doesn't warrant a direct comparison. Rather, we are interested in comparing models' performance on different structures and their trend to the human trend.

As seen from the table, both ruGPT and ruBERT perform very well. Moreover, they show similar error patterns to humans. The sentences where it was easier to distinguish the correct sentence from an incorrect one were sentences where both the subject and the attractor were of the same number, especially in the singular. Humans show better results on completing such sentences as well. However, when the subject and the attractor differ in number, for humans it was easier when the subject was in plural, while for both ruBERT and ruGPT this was more difficult and they made less mistakes in singular subject + plural attractor structure.

4.2 Comparison of attention scores

We compare attention scores between sentences of different structures. We calculate paired Student's test; Figure 2 shows p-values of such tests after Bonferroni correction for ruBERT and ruGPT respectively. As seen from the figure, for ruBERT model, the main significant differences ($p < 0.05$) are mostly between correct sentences and similar sentences with attractors. For example, a correct sentence of type P_P-P (predicate, subject and attractor are in plural) is significantly different from structures P_S-P (predicate in plural, subject in singular and attractor in plural) and S_P-S (predicate in singular, subject in plural and attractor in singular). However, grammatical sentences do not differ in attention with ungrammatical sentences

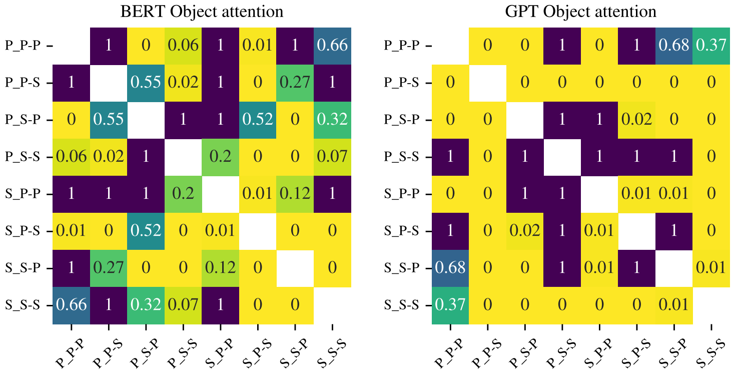


Figure 2: Results of Student's pairwise t-test (p-values) for ruBERT and ruGPT verb-to-attractor attention scores respectively between 8 variants of the sentence. The first letter encodes the number of a predicate (*S* for singular and *P* for plural), the second letter encodes the number of a subject and the third letter encodes the number of an attractor, for example, *P_S-P* stands for a sentence where a subject is in singular but an attractor and a predicate are in plural.

where attractors differ in number with predicates.

Attention scores in ruGPT do not follow a clear pattern and are most probably affected by other factors that we do not control as we focus on difference in number.

Structure	**ruGPT**	**ruBERT**	**Humans**
S-S	**1.0**	**1.0**	**0.83**
S-P	**1.0**	0.94	0.77
P-S	0.95	0.86	0.79
P-P	**1.0**	0.97	0.8

Table 1: Comparison of accuracy *scores* of ruGPT and ruBERT to the *percentage* of successfully completed tasks in the psycholinguistic experiment (figures for humans are taken from Slioussar (2018)

4.3 Comparison with human results

We employ statistical models similar to those of Slioussar (2018) and perform regression analysis in R (R Core Team, 2023) with mixed models from *lme4* package (Bates et al., 2015). We employed package *lmerTest* Kuznetsova et al. (2017), and also *pbkrtest* Halekoh and Højsgaard (2014), where applicable, to obtain the p-values of variables in these mixed models. Below we report p-value of *lmerTest* but Kenwald-Roger test of *pbkrtest* yields very similar p-values numerically. The R code for these calculations is also available in our project repository.

The following comparison is made to data on human reading times (RT) (Slioussar, 2018), on which we fit a new model. We evaluate the performance of both language models with the following mixed-effects statistical model (7). The dependent variable is the score (or RT in humans) of a singular and of a plural predicate given a certain subject-attractor prefix. Recall, that for every sentence, there are 4 possible prefixes and 2 possible numbers for the predicate, thus we have 8 sentence variants. This is a setup similar to Slioussar experiments with humans' RTs when reading such sentences word-by-word. We thus compare RT for humans with scores for our models. Although these are, of course, quite disparate values, we deem them to be the most optimal values for comparison in the available data. These are both numeric variables, which we take to be proxying 'surprisal' by a given sentence.

Slioussar shows that RTs are, in a sense, delayed and that predictor variables (described below) are not significant on the word 4, the verb, first word of the predicate, but significant on word 5, the participle, second word of the predicate. Our model fitted on word 4 is indeed not significant, thus for humans we analyze RTs on word 5. We reiterate that for models we test verb/word 4 scores. Models and humans are different in how they process sentences, and we consider such setup to be a fair comparison.

(7) $\mathrm{lmer}(\mathrm{Score} \sim N_1 + N_2 + \mathrm{kind} + (1|\mathrm{Sent}))$

As random effects we use sentence number (and participant number for humans, too). Our predictor variables are the number of the sub-

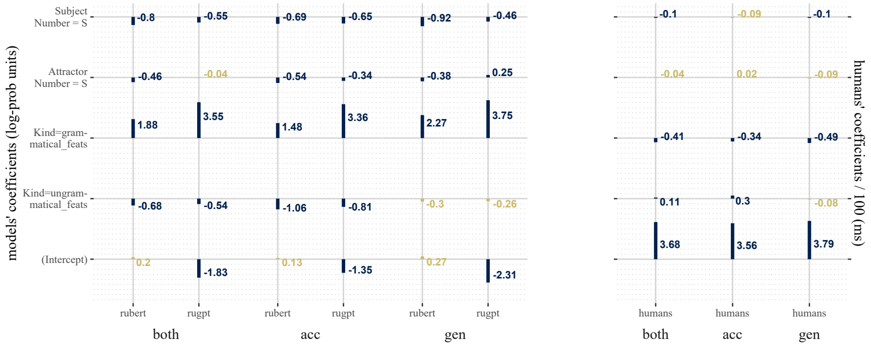


Figure 3: Estimates of variables predicting score for models (3.3.1), and reaction time (ms) for humans in data from (Slioussar, 2018). These proxy 'surprisal' but differently: *lower model score* → more surprisal, *higher human reaction time* → more surprisal. Coefficients are in **dark blue**, when p-values computed with *lmerTest* package (Kuznetsova et al., 2017) are significant with p-value $< \alpha = 0.05$.

ject and of the attractor, similarly to Slioussar (2018), but unlike Slioussar we do not include interaction terms into the model, opting instead for three-way encoding of sentence 'grammaticality' judged by *grammatical features*. We distinguish *gram*matical sentences, where verb number matches subject number, *attr*actor sentences, where verb number does not match subject number but matches attractor number and *ungram*matical sentences, where verb number matches neither. We encode contrasts, such that $kind = attractor$ falls into intercept, and $kind = grammatical$ and $kind = ungrammatical$ remain as (one-hot encoded) variables. Thus, while we also test attraction effects as does Slioussar, our approach allows us to test the ranking hypothesis. Recall, that we predict the ranking in (6) for 'surprisal'. For models this should be exactly the ranking of scores and for humans this should be the reverse ranking of RTs (least time spent on grammatical sentences and most on ungrammatical). We show below, that this is mostly borne out.

Finally, recall that sentences of Slioussar (2018) each belong to one of the two types: in half of the sentences the attractor is in accusative case and in the other half in the genitive case, and case determines surface syncretism (Section 3.2). Thus syncretism is captured differently: for accusative, where ACC.PL = NOM.PL – by 'kind' variable above, for genitive, where GEN.SG = NOM.PL – by 'attractor number' variable.

We thus perform regression analysis analysis for three sets of data: all sentences, accusative case only sentences, genitive case only sentences. The first model would inform us of grammatical tendencies, while the other two models isolating case would inform us of the effect of surface syncretism. These three regressions are fit on each of ruBERT, ruGPT and humans data, totalling in 9 experiments.

P-values and coefficients are shown in Figure 3. On full data, for ruBERT all variables achieve significance and for ruGPT all variables except attractor number achieve significance. This correlates with investigation into attention heads: there, similarly, ruBERT seems to attend to the attractor, while ruGPT does not. The ranking hypothesis in 6 holds, and grammatical sentences receive higher scores (W > 0) than baseline, attractor sentences, while absolutely bad sentences (non-grammatical and without attraction) receive scores lower (W < 0) than baseline attractor sentences. Importantly, for ruGPT the bigger coefficients indicate stronger distinction between sentence kinds. This is in line with its higher accuracy (Table 1). As for the human data, the result is similar to ruGPT, rather than ruBERT, with attractor number not achieving significance. However, the ranking holds for humans too: RTs to grammatical sentences are lower (W < 0) than for attractor sentences (interpreted as less surprisal) while they are higher (W > 0) for totally ungrammatical sentences.

We now consider two subsets by case independently. For accusative case sentences, where syncretism is exactly the (PL = NOM.PL) the results are very similar to full data results. The ranking hypothesis holds. This is because exactly this type of syncretism (deep, grammatical) is captured well by feature *kind* variable. As such, for ruBERT and ruGPT all variables achieve significance, even attractor number for ruGPT. Again, GPT coefficients are higher indicating stronger distinction between sentence kinds. As for human RTs, they are significantly different between sentence kinds and follow the hypothesis. However, the precise numbers of subject and attractor are insignificant, which would mean there is no assymetry in agreement with singular or plural nouns for humans.

Finally, we consider only the sentences with genitive, where syncretism is GEN.SG=NOM.PL, so if surface form matters in attraction, singular would be more "attractive" here than grammatical plural. This is not captured by *kind* variable, which is oriented on grammatical feature rather than surface form. Thus featurally ungrammatical sentences are not significantly different from feature attracting sentences. However, the attractor number being singular increases the score of ruGPT, while for full data and accusative data the reverse was true. Put more explicitly, it means that grammatical attraction in genitive plural (*P_S-P*: *Subject* =S, *kind=attractor*, *Attractor*=P $\implies$ *Attractor*=*Verb*=PL) is scored at $-2.31 + (-0.46) = -2.77$, *lower* than surface form attraction in genitive singular (*S_S-P*, techinically *kind=ungram*, *Subject* = *Attractor* =S $\implies$ *Verb*=PL) $-2.31 + (-0.46) + 0.25 = -2.52$ (not counting the insignificant *kind=ungram* = -0.26). Although ruBERT result is inconclusive (perhaps due to intercept not being significant) we take this to indicate that at least for GPT it is formal syncretism and not grammatical features, that predicts the attraction. This is a result similar to (Slioussar, 2018). Our model on her human data shows similar result: RT is not significantly different between featurally "attractive" and ungrammatical sentences, while singular attractor reduces RT.

Overall, models seem more sensitive to attractor number than humans, meaning singular and plural attractors are treated differently in a setup where attraction by grammatical number could happen.

5 Conclusion

We explored how models react to errors in subject-verb agreement, where humans are prone to mistakes of *attraction*. These are ungrammatical contexts that look as if agreement happens not with the subject as a whole, but with subject's dependent (2a, 33a).

We find that indeed, like humans, models see such sentences as more acceptable than ungrammatical sentences with no attraction, i.e. the ranking in (6) holds for humans and models alike. Most importantly, we find in genitive, a pattern similar to what Slioussar (2018) finds, where surface syncretism is more predictive of attraction than grammatical number. Recall that in our case attraction by surface syncretism obtains for genitive singular, where the attractor is neither nominative nor plural, while grammatical attraction is expected for genitive plural. This is a somewhat puzzling result for humans (Slioussar, 2018) and models alike, because other tasks show that both are sensitive to deeper structure.

As for overall accuracy, ruGPT, a decoder model, was more likely to choose correct sentence continuation, assigning higher probability to the verb form with the correct number. BERT, an encoder model, did worse here.

Attention scores investigation does not present a clear picture, but for ruBERT comparison between sentences that differ only in attractor are significant. This may be the reason for why its scores are significantly determined by attractor number.

We examined a single phenomenon of agreement attraction in subject-verb agreement on a constrained dataset from a psycholinguistic study of Slioussar (2018). We confirmed that ruBERT and ruGPT exhibit agreement attraction by grammatical number. An intriguing preliminary finding, resembling Slioussar (2018)'s results is that for ruGPT agreement attraction seems more sensitive to formal identity than to grammatical number, which could be distinguished in Russian genitive forms.

6 Limitations

This study presents several limitations that necessitate further investigation. The study's findings are based on a single experiment focusing on grammatical number agreement and only on one language. Moreover, a single and frequent verb lemma is tested. This narrow scope limits the generalizability of the results to other grammatical phenomena.

Future research should explore the observed effects across a wider range of grammatical structures.

The study compared human performance to that of language models based on the assumption that these models demonstrate sensitivity to probabilistic relationships at the word level. However, this comparison remains indirect. Although the selected models allowed direct comparisons under specific experimental conditions and successfully reproduced previously observed grammaticality effects, other models, even within the same architecture may show different results. Future research would benefit from exploring the nuances of different language model architectures in relation to human performance in grammaticality tasks.

Additionally, large language models are used for research, which implies that even inference on such models can be difficult with a limited computational budget.

7 Ethics Statement

In the implementation and evaluation of our proposed approach, we use only publicly available code to avoid any ethical concerns. We use data acquired upon request (to Slioussar). The data did not include any personal data, as each participant was encoded with a label. i.e. participant 1, 2 etc. To the best of our knowledge, all participants gave an informed consent to the author of original studies.

References

Suhas Arehalli and Tal Linzen. 2020. Neural language models capture some, but not all, agreement attraction effects.

M. Baerman, D. Brown, and G.G. Corbett. 2005. *The Syntax-Morphology Interface: A Study of Syncretism*. Cambridge Studies in Linguistics. Cambridge University Press.

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.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural nlp. In *Proceedings of the 58th annual meeting of the association for computational linguistics: tutorial abstracts*, pages 1–5.

Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi, editors. 2023. *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP (Organizing committee message)*. Association for Computational Linguistics, Singapore.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Pavel Caha. 2019. Syncretism in morphology.

Noam Chomsky. 1965. Aspects of the theory of syntax cambridge. *Multilingual Matters: MIT Press*.

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

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.

Maria Teresa Guasti. 2017. *Language acquisition: The growth of grammar*. MIT press.

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.

Ulrich Halekoh and Søren Højsgaard. 2014. A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models – the R package pbkrtest. *Journal of Statistical Software*, 59(9):1–30.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Carina Kauf and Anna Ivanova. 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 2: Short Papers), pages 925–935, Toronto, Canada. Association for Computational Linguistics.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

Andrew Kyle Lampinen. 2022. Can language models handle recursively nested grammatical structures? a case study on comparing models and humans. *arXiv preprint arXiv:2210.15303*.

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.

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.

Bai Li, Zining Zhu, Guillaume Thomas, Yang Xu, and Frank Rudzicz. 2021. How is BERT surprised? layerwise detection of linguistic anomalies. 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 4215–4228, Online. Association for Computational Linguistics.

Tal Linzen and Brian Leonard. 2018. Distinct patterns of syntactic agreement errors in recurrent networks and humans.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Carson T. Schütze. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Natalia Slioussar. 2018. Forms and features: The role of syncretism in number agreement attraction. *Journal of Memory and Language*, 101:51–63.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. Rublimp: Russian benchmark of linguistic minimal pairs.

William Timkey and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8705–8720, Singapore. Association for Computational Linguistics.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Matthew W. Wagers, Ellen F. Lau, and Colin Phillips. 2009. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2):206–237.

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.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

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.

Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergei S. Markov, Vladislav Mikhailov, and Alena Fenogenova. 2024. A family of pretrained transformer language models for Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 507–524, Torino, Italia. ELRA and ICCL.