

Subject Verb Agreement Error Patterns in Meaningless Sentences: Humans vs. BERT

Karim Lasri^{α,β} Olga Seminck^α Alessandro Lenci^β Thierry Poibeau^α

^αLattice (École Normale Supérieure-PSL, CNRS, U. Sorbonne Nouvelle)

^βUniversity of Pisa

karim.lasri@ens.psl.eu olga.seminck@cnrs.fr
alessandro.lenci@unipi.it thierry.poibeau@ens.psl.eu

Abstract

Both humans and neural language models are able to perform subject-verb number agreement (SVA). In principle, semantics shouldn't interfere with this task, which only requires syntactic knowledge. In this work we test whether meaning interferes with this type of agreement in English in syntactic structures of various complexities. To do so, we generate both semantically well-formed and nonsensical items. We compare the performance of BERT-base to that of humans, obtained with a psycholinguistic online crowdsourcing experiment. We find that BERT and humans are both sensitive to our semantic manipulation: They fail more often when presented with nonsensical items, especially when their syntactic structure features an attractor (a noun phrase between the subject and the verb that has not the same number as the subject). We also find that the effect of meaningfulness on SVA errors is stronger for BERT than for humans, showing higher lexical sensitivity of the former on this task.

1 Introduction

Subject Verb Agreement (SVA) is a grammatical constraint in English and several other natural languages, such that verbs must agree in number with their subject. Linguistic theories generally assume that SVA obeys two main principles: i.) **structure dependence** (SD) - SVA is governed by phrasal structure, rather than surface linear order (i.e., the verb agrees with the syntactic subject); ii.) **meaning independence** (MI) - SVA is a morphosyntactic constraint that holds for meaningless sentences too (e.g., *Colorless green ideas sleep furiously*) (Chomsky, 1956, 1971; Chomsky et al., 1976). However, previous research has shown that humans are prone to making agreement errors with specific constructions (Bock and Miller, 1991; Hartsuiker et al., 2001), for example when a noun with a different number (also called an **attractor**) occurs between the subject (the **cue**) and the verb (the **tar-**

get). See (1) from Bock and Miller (1991) for an example where agreement can be disturbed by an attractor, as human subjects often show preference for a syntactically ill-formed sentence:

- (1) [The **readiness**]_{subject} [of our conventional **forces**]_{attractor}]_{PP} [**are**]_{verb} at an all-time low.

This evidence suggests that the SD principle of SVA might be weaker than it is typically assumed and can be disrupted or disturbed under specific conditions. At the same time, such violation prompts the need to carefully test whether the MI principle of SVA is also compromised in subjects' grammaticality judgments.

SVA has become a widespread testbed to investigate the syntactic knowledge that **neural language models** (NLMs) are able to acquire. A key question is to explore to what extent their competence of syntax obeys the same constraints as those of humans, by comparing the behavior of NLMs with subjects' judgments. In this paper, we pursue this goal by focusing on the SD and MI properties of SVA in humans and NLMs. Its contribution is twofold: i.) we collect original human data on meaningful and meaningless stimuli featuring syntactic structures of varying complexities; ii.) we analyze and compare the error patterns in humans and in NLMs. This allows us to address the following questions: i.) are the SD and MI assumptions about SVA truly supported by human judgments? ii.) do humans and NLMs make similar SVA errors in structures with attractors and/or in meaningless sentences? This may help understand to what extent NLMs rely on syntactic knowledge abstracted from training examples. By comparing the error patterns of humans and NLMs on SVA, we can derive important information about the nature of their linguistic competence: Is the ability of NLMs on SVA completely meaning independent? Is it influenced by the complexity of the syntactic structure?

Structure	Structure description	Example
A	Simple agreement	[The author] _{subject} [<i>laughs/*laugh</i>] _{verb}
B	In a sentential complement	[The mechanics] _{subject} [said] _{vComp} [[the author] _{SComp} [<i>laughs/*laugh</i>] _{verb}] _{Comp}
C	Across a prepositional phrase	[The mechanic] _{subject} [near the author] _{pp} [<i>smiles/*smile</i>] _{verb}
D	Across a subject relative clause	[The author] _{subject} [that likes the movie] _{subj. RC} [<i>laughs/*laugh</i>] _{verb}
E	In a short verb phrase coordination	[The author] _{subject} [laughs and <i>swims/*swim</i>] _{VP Coord}
F	Across an object relative clause	[The author] _{subject} [that the mechanics like] _{obj. RC} [<i>smiles/*smile</i>] _{verb}
G	Within an object relative clause	[The author] _{subject} [that the mechanics like/*likes] _{obj. RC} [smiles] _{verb}
H	Across an object relative clause (<i>no that</i>)	[The author] _{subject} [the mechanics like] _{obj. RC} [<i>smiles/*smile</i>] _{verb}
I	Within an object relative clause (<i>no that</i>)	[The author] _{subject} [the mechanics like/*likes] _{obj. RC} [smiles] _{verb}

Table 1: Agreement structures used in this study, from [Marvin and Linzen \(2018\)](#). The cue is bolded, the target is in italic, and the attractor is underlined. For each target, we display both the correct and incorrect inflections.

2 Related Work

[Linzen et al. \(2016\)](#) tested Long Short Term Memory (LSTM) language models and their performance on the SVA task. They found that these models can capture a non-trivial amount of grammatical structure but that they are insufficient for capturing complex syntax-sensitive dependencies. In [Marvin and Linzen \(2018\)](#) the capacity of LSTM language models to perform the SVA task is compared with human data. In particular for sentences including attractors, the models perform worse than humans. [Linzen and Leonard \(2018\)](#) tested recurrent neural networks (RNNs) on the task to see whether the error pattern was similar to humans. They concluded that despite important similarities, there was a different behavior within relative clauses. In particular, RNNs are sensitive to the number of attractors, whereas humans are not.

[Gulordava et al. \(2018\)](#) investigated the role of semantics on the performance of RNNs on SVA, testing such neural models against meaningless, or “nonce”, sentences built from various syntactic constructions. They used meaningless sentences where RNNs could not rely on semantic or lexical cues. For Italian, they found that the performance of an LSTM model and the performance of humans were comparable for meaningless items and semantically congruent ones. [Lasri et al. \(2022\)](#) tested a transformer model, BERT-base ([Devlin et al., 2019](#)), on its capacity of subject-verb agreement on the items from [Marvin and Linzen \(2018\)](#) and on semantically disrupted sentences featuring the same syntactic constructions. They found that although the model generalized well for simple templates, it failed on meaningless items with attractors.

3 Experimental Setup

In this section, we describe the procedure to construct the experimental items used to collect human

judgments with crowdsourcing and to test NLM behavior on SVA.

Items We test the syntactic structures in Table 1, four of which include an attractor. For each syntactic template, we generate 30 meaningless sentences by sampling random words with the correct part-of-speech at each position of the template, using an extensive vocabulary (see Appendix A for more details). The generated sentences are meaningless in the sense that our generation procedure does not include any semantic constraint designed to grant meaning to the sentence. The resulting word sequence should therefore be semantically ill-formed, similar to the classic example *Colorless green ideas sleep furiously*. As some generated sentences can accidentally happen to be interpretable, we also tried to manually remove such sentences when crafting our stimuli. However, as meaning is a subtle notion, some sentences might be interpretable in very specific contexts after our manual filtering. Still, we expect such cases to be very rare in our dataset. Every minimal pair consists of sentences similar to (2).

- (2) a. *The admissions sings.
b. The admissions sing.

We also sample 30 meaningful sentences for each structure from [Marvin and Linzen \(2018\)](#) to collect human data.¹ As the filtered M&L data used to collect human performances resulted in a very limited number of sentences built on a limited vocabulary, we extract BERT’s performance on the whole dataset for M&L, following the procedure in [Goldberg \(2019\)](#).

We thus collect human performance on 30 items for each of our 2 conditions (Nonce and M&L), and each of our 9 syntactic structures, for a total of

¹We will call this data **M&L**. We filter out sentences where the verb is ‘be’, as this verb is very frequent in English.

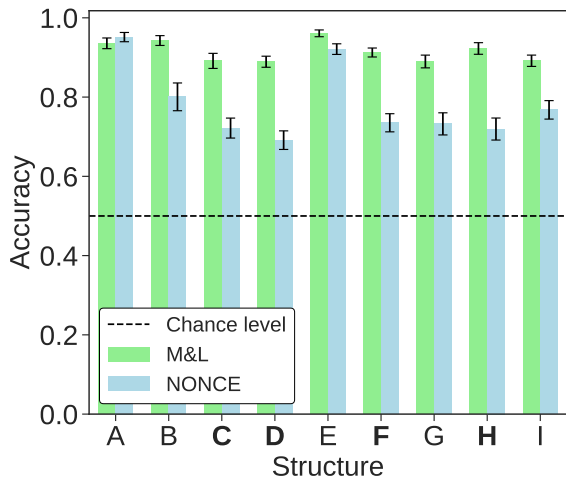


Figure 1: Human accuracies on the SVA task. Structures where an attractor is present are displayed in bold. The error bars displayed represent the 95% confidence interval.

540 items. More details are given in Appendix B.²

Neural Language Model Our tested NLM is BERT (Devlin et al., 2019), a bi-directional pre-trained transformer model. BERT has been shown to possess a number of syntactic abilities (Jawahar et al., 2019; Goldberg, 2019), the nature of which remain scarcely understood (Baroni, 2019). For each item, we present BERT with sentences masked at the target position and compare the probability that BERT assigns to each verb inflection.

Collection of Human Judgements We collect our human data using the online click working platform Prolific.³ We implemented a binary choice experiment in Psychopy (Peirce, 2007) hosted on Pavlovia⁴ where participants were presented with a minimal pair, such as in (2), and asked which sentence was the most correct. In order to prevent habituation to our stimuli and task, we used 64% of filler items. We recruited 300 participants to obtain 20 responses per item and kept the responses of 270 participants in our final data set. Their mean speed to judge one item was 6.9 seconds. See Appendix B for more details about human data collection.

²Our data is available at <https://github.com/karimlasri/agreement-humans-bert>

³<http://prolific.co>

⁴<https://pavlovia.org>

4 Results

We first discuss the human data, which we then compare with BERT’s performance.

4.1 Error Patterns for Humans

In this analysis, we compare human accuracy on the nonce stimuli and on Marvin and Linzen’s (2018) sentences. In Figure 1, we break down the results by syntactic structure to observe whether the construction type affects the human judgments. We notice a performance drop in all structures with nonce sentences, except for A where the apparent increase is not significant, as shown by the error bars. Interestingly, the structures with an attractor (bolded in the x -axis) are those for which the performance drop seems to be the highest. We also observe high performance drops in sentences where there is no attractor (B, G and I).

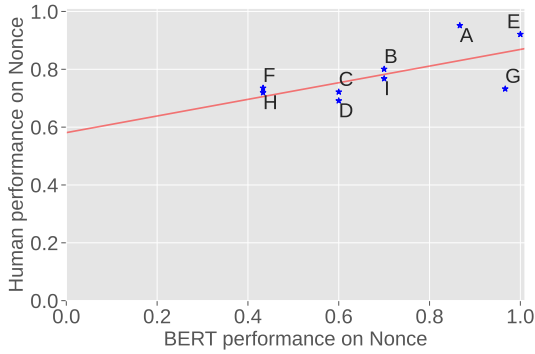
Looking at Table 1, we can see that these structures are more complex than the structures where the effect of meaningfulness is low (A and E). Indeed, they contain either a complement (B), or a relative clause (G, I). Surprisingly, we observe a similar pattern on meaningless sentences in (F) and (G): humans seem to be perturbed as much by the attractor within the object relative clause (F), as they are by the material in the main clause (G), if sentences are meaningless. This evidence in comprehension seems opposite to Bock and Cutting’s (1992) claim that agreement production is only sensitive to information within the clause of the target. This evidence hints at the possibility of a difference in the mechanisms that support SVA in production and comprehension.

Metric	Correlation
M&L Accuracy	0.61
Nonce Accuracy	0.65
Accuracy Drop	0.52

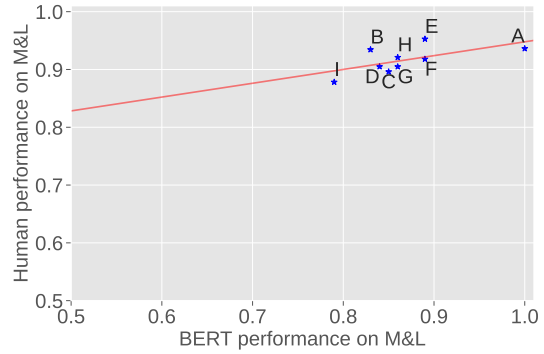
Table 2: Coefficient of determination between BERT’s and human performance on SVA, averaged across syntactic structures. The accuracy drop condition represents the difference between average performance on M&L’s stimuli and our nonce stimuli, as seen in Figure 3.

4.2 How Similar are the Human and BERT’s Error Patterns?

In this analysis, we compare the performance achieved by BERT with human performance, on each of our stimuli types. Figure 2 displays the



(a) Human nonce performance vs. BERT



(b) Human M&L performance vs. BERT

Figure 2: A comparison of human performance against BERT’s performance on each of our structures.

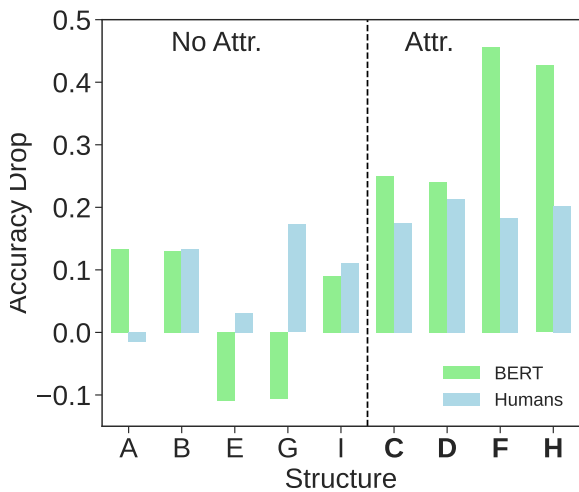


Figure 3: Performance drops between M&L and nonce stimuli.

result obtained by humans against BERT’s performance for each syntactic template, for both meaningful and meaningless sentences. Interestingly, there seems to be a fairly high alignment between the results for each syntactic construction, and for each source of stimuli. We display the R^2 correlation measurement of our fit in Table 2. The latter confirms the observed alignment, as we obtain quite high correlations (0.61 for meaningful sentences and 0.65 for nonce sentences).

However, we observe that while the variation in performance obtained by humans across templates seems quite low, BERT’s performance does seem to be more affected by the different structures. This is especially true in the case of nonce sentences, as seen in Figure 2a. We also observe a difference in performance decrease in Figure 3, as BERT’s performance drops are overall higher in presence of an attractor compared to those of humans. On

the other hand, BERT has a higher performance drop on (A) and humans on (G). This in turn could be explained by the fact that (G) is a hard sentence to process for humans, the target of the agreement being within an embedded relative clause, while BERT could rely on local context in this case as observed by Lasri et al. (2022).

5 Discussion

5.1 Lexicalization and Syntactic Generalization

While subject-verb agreement is sometimes considered as a purely syntactic phenomenon, our results show that actually humans also rely on semantics, which goes against the meaning independence hypothesis. Our results also show that BERT is also highly dependent on semantics, a finding in line with Bernardy and Lappin (2017), who mention that “*deep neural networks require large vocabularies to form substantive lexical embeddings in order to learn structural patterns*”. This highlights the strong connection between the ability to process linguistic structure and the semantic content of sentences.

5.2 Structure Dependence

Throughout this study, we observed that the performance of both humans and BERT were sensitive to the syntactic structure used in our items. Humans clearly obtain lower performance on sentences that are more complex to process when they are meaningless, including but not limited to sentences presenting an attractor. This variation in performance seems to reflect variation in structure complexity, which upholds SD. On the other hand, BERT seems to be mostly sensitive to sentences with attractors. This evidence rather shows a violation of SD, as at-

tractors are only related to the target by linear order, in line with evidence found by [Lasri et al. \(2022\)](#). While human and BERT’s results seem to correlate to a large extent, these divergences could reflect a difference in processing. For instance, SVA in sentence comprehension for humans could depend on having read the whole sentence, while BERT could rely more on local context for this task. Indeed, a fine-grained analysis performed in previous work showed BERT to be mostly sensitive to the replacement of linearly close tokens ([Lasri et al., 2022](#)).

6 Conclusion

Throughout this work, we have shown that the ability to perform SVA is highly dependent on the syntactic construction when presented with meaningless sentences. The failures of humans seem to align well with those of BERT overall, as sentences with attractors tend to compromise meaning independence when processing the agreement relation. Despite these similarities, we further show that the performance drop is generally higher in BERT on meaningless sentences, and that humans are more perturbed by complex constructions without an attractor. This finding can in turn reflect differences in processing syntactic structure, either reflecting more reliance on local context for BERT, or a difference between agreement processing in production and comprehension, which could be the source of the partial mismatch in the observed error patterns.

Ethics Statement

The authors foresee no ethical concerns with the work presented in this paper.

Acknowledgements

This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

Marco Baroni. 2019. [Linguistic generalization and compositionality in modern artificial neural networks](#). *Philosophical Transactions of the Royal Society B*, 375(1791).

Jean-Phillipe Bernardy and Shalom Lappin. 2017. [Using deep neural networks to learn syntactic agree-](#)

[ment](#). In *Linguistic Issues in Language Technology, Volume 15*, 2017. CSLI Publications.

- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O’Reilly Media, Inc."
- Kathryn Bock and J.Cooper Cutting. 1992. [Regulating mental energy: Performance units in language production](#). *Journal of Memory and Language*, 31(1):99–127.
- Kathryn Bock and Carol A Miller. 1991. Broken agreement. *Cognitive psychology*, 23(1):45–93.
- Noam Chomsky. 1956. Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124.
- Noam Chomsky. 1971. *Problems of knowledge and freedom: The Russell lectures*. Pantheon Books.
- Noam Chomsky et al. 1976. *Reflections on language*. Temple Smith London.
- European Commission. 2018. [2018 reform of eu data protection rules](#).
- 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.
- Yoav Goldberg. 2019. [Assessing bert’s syntactic abilities](#). *CoRR*, abs/1901.05287.
- Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. [Colorless green recurrent networks dream hierarchically](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.
- Robert J Hartsuiker, Inés Antón-Méndez, and Marije Van Zee. 2001. [Object attraction in subject-verb agreement construction](#). *Journal of Memory and Language*, 45(4):546–572.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. [What does BERT learn about the structure of language?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Karim Lasri, Alessandro Lenci, and Thierry Poibeau. 2022. [Does BERT really agree ? fine-grained analysis of lexical dependence on a syntactic task](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2309–2315, Dublin, Ireland. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. [Assessing the ability of LSTMs to learn syntax-sensitive dependencies](#). *Transactions of the Association for Computational Linguistics*, 4:521–535.

Tal Linzen and Brian Leonard. 2018. [Distinct patterns of syntactic agreement errors in recurrent networks and humans](#). *CoRR*, abs/1807.06882.

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*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Jonathan W Peirce. 2007. Psychopy—psychophysics software in python. *Journal of neuroscience methods*, 162(1-2):8–13.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. [Luminosoinight/wordfreq:v2.2](#).

A Vocabulary Selection

We used the vocabulary of [Lasri et al. \(2022\)](#) for the generation of our nonce items and filtered it. We selected a vocabulary of nouns and verbs that checked the following criteria:

1. We filter out tokens that are ambiguous, i.e. tokens which can either be a noun or a verb. We used Wordnet ([Miller, 1995](#)) implemented in the NLTK library ([Bird et al., 2009](#)) in python 3 to check whether a word was not classified as a noun and a verb by checking whether there was no synset in the other category.
2. We filter out words using their relative frequency measured by using the python library wordfreq ([Speer et al., 2018](#)).⁵ We choose to filter too frequent words because some were ambiguous with another category (e.g. in the noun vocabulary we can find *good, well, one*). We decided to remove infrequent words to prevent that participants would not know their meaning. For example, this filtered out *polynomial*, and *consonant* from the noun vocabulary.

⁵<https://pypi.org/project/wordfreq/>

3. We only keep words with a length ranging from 3 and 8 characters, in order to prevent big differences in size between the items produced by one template.

4. We make a subdivision between transitive and intransitive verbs in order to correctly fill the templates.

B Collection of Human Judgments

Setup To collect our human judgements, we recruited participants on the working platform ProLific. Participants were redirected to the Pavlovia page hosting our experiment. First, they had to give their informed consent. Their data was processed in accordance to the European General Data Protection Regulation ([Commission, 2018](#)), and no sensitive data has been collected. After being informed, participants were shown brief instructions about the forced choice task. For each item, they were presented with two sentences, and asked to select the one that seemed more acceptable using the keyboard arrows. Each session started with three training items followed by feedback. When the training was finished, they were notified that the experiment started and that they would not receive feedback anymore. Each participant was presented with 100 items and thereafter received a message that the experiment was over.

Number of Items and Participants In total, the participants replied to 100 items: 64 fillers, and 36 experimental item. 18 were from the nonsense condition and 18 from the M&L data set. As every condition features 9 different structures, 2 structures of each category where shown per participant. In order to collect 20 responses per item, we recruited 300 participants with 15 different versions of the online experiment, which can be found in the supplement material of this article.

Fillers Our filler items where from [Ettinger \(2020\)](#). They feature semantically appropriate and inappropriate completions. We also used filler items with correct and incorrect determiners among ‘a/an’ depending on the following noun to feature syntax-oriented fillers as well.

Selection of Participants We only accepted participants with the United States nationality with English as a first language between the age 18 and 60 years old. We excluded participants that already contributed to another version of the experiment.

Reward Participants got rewarded 2.25£ for a participation that was estimated to take 15 minutes. This was estimated to be a ‘good’ hourly pay by the Prolific platform. We rejected participants that performed the experiment in less than 5 minutes.

Filtering and Loss of Participants In our final data set we have the contributions of 270 participants. Participants that performed at chance level (50 % accuracy) were filtered out. Furthermore, we lost some data of participants that did not close the experiment correctly in Pavlovia.