Assessing the influence of attractor-verb distance on grammatical agreement in humans and language models

Christos-Nikolaos Zacharopoulos
Cognitive Neuroimaging Unit, NeuroSpin center, France
Sensome SAS, Massy, France
christonik@gmail.com

Théo Desbordes
Meta AI Research;
Cognitive Neuroimaging Unit, NeuroSpin center, France;
desbordes.theo@gmail.com

Mathias Sablé-Meyer
Cognitive Neuroimaging Unit, NeuroSpin center, France;
Collège de France, Université PSL
mathias.sable-meyer@ens-cachan.fr

Abstract

Subject-verb agreement in the presence of an attractor noun located between the main noun and the verb elicits complex behavior: judgments of grammaticality are modulated by the grammatical features of the attractor. For example, in the sentence *"The girl near the boys likes climbing"*, the attractor (*boys*) disagrees in grammatical number with the verb (*likes*), creating a locally implausible transition probability. Here, we parametrically modulate the distance between the attractor and the verb while keeping the length of the sentence equal. We evaluate the performance of both humans and two artificial neural network models: both make more mistakes when the attractor is closer to the verb, but neural networks get close to the chance level while humans are mostly able to overcome the attractor interference. Additionally, we report a linear effect of attractor distance on reaction times. We hypothesize that a possible reason for the proximity effect is the calculation of transition probabilities between adjacent words. Nevertheless, classical models of attraction such as the cue-based model might suffice to explain this phenomenon, thus paving the way for new research. Data and analyses available at https://osf.io/d4g6k

1 Introduction

On the surface, language appears to be produced and understood linearly, as humans read or hear words one after the other. Yet, formal linguistic theories postulate the existence of an underlying structure that governs language processing (Chomsky, 1957; Rizzi, 2004; Dehaene et al., 2015; Vigliocco and Nicol, 1998; Hauser et al., 2002). This hypothesis is supported by both behavioral (Fossum and Levy, 2012; Shi et al., 2020; Coopmans et al., 2021) and neural (Brennan et al.; Nelson et al., 2017; Pallier et al., 2011) data. According to a competing view, unstructured probabilistic models capture behavior without explicitly relying on structures (Frank and Bod, 2011). The discrepancy between the two views has led to a decade-long debate on linear versus structural effects in language (Haskell and MacDonald, 2005; Ding et al., 2015; Willer Gold et al., 2017; Arana et al., 2021).

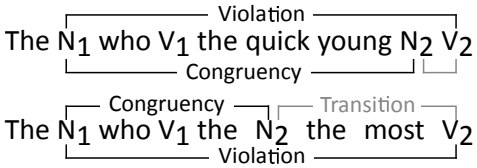

Figure 1: **Experimental design** Our design aims to separate two possible mechanisms involved in language processing, by altering the distance of the embedded noun (attractor) from the verb. We compare the impact of noun-verb agreement (*Violation*) and noun-noun dependency (*Congruency*). The *Transition* effect symbolizes the consequence of the transition probability between the second noun and the verb. As a baseline, we also include a condition with a similar number of words, but without a noun in between. The summary of the experimental conditions is shown in Table 1

Analyzing subject-verb agreement errors in the presence of distracting elements is a standard way of separating linear and structural accounts of language (Molinaro et al., 2011). Attraction effects in number agreement have been studied extensively in humans (Bock and Miller, 1991; Franck et al., 2002; Hammerly et al., 2019) and Neural Language Models (NLMs) (Linzen et al., 2016; Jumelet et al., 2019; Lakretz et al., 2021).

In the present work, we introduce a parametric manipulation of the distance between the distracting nouns and the verb (Figure 1). We include a baseline condition where the subject-verb distance is matched but no word that carries number marking is introduced in-between. We thereby provide a minimal triplet of experimental conditions (Table 1) that can disentangle structural from linear mechanisms by contrasting operations directly ascribed to each mechanism. We posit that there are structural operations at play to explain the fact that participants are always able to detect violations,

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 16081–16090

Condition	Sentence	Violation	Congruency
Proximal Attractor	The writer who knows the happy young journalist climbs.	✗	✓
	The writer who knows the happy young journalists climbs.	✗	✗
	The writer who knows the happy young journalist climb.	✓	✓
	The writer who knows the happy young journalists climb.	✓	✗
Distal Attractor	The writer who knows the journalist the most climbs.	✗	✓
	The writer who knows the journalists the most climbs.	✗	✗
	The writer who knows the journalist the most climb.	✓	✓
	The writer who knows the journalists the most climb.	✓	✗
Baseline	The writer who walks fast but rather clumsily climbs.	✗	-
	The writer who walks fast but rather clumsily climb.	✓	-
Filler (Number)	The writer who knows the journalist climbs.	✗	-
	The writer who know the journalist climbs.	✓	-
Filler (POS)	The writer whom the journalist knows climbs.	✗	-
	The writer whom knows the journalist climbs.	✓	-

Table 1: **Conditions.** The design contrasts the attribution of two distinct effects to the overall error rate in humans and NLMs. To prevent the subjects from developing strategies for the effective resolution of the violation-detection task, we included two different kinds of filler trials that contained violations at the beginning of the sentence.

and at the same time there are linear operations at play which lead to the attraction effect of the intervening noun. We additionally hypothesize that the linear effect might be modulated by transition probabilities, i.e., the prior probability that a given word follows another (Dehaene et al., 2015; Friston et al., 2021). Finally, despite reaching on par with or even supra-human performance on many tasks (Brown et al., 2020; Minaee et al., 2021), NLMs are known to be sensitive to superficial statistical properties of their training data (McCoy et al., 2021). Thus, we compare two NLMs with the behavior of human participants.

2 Experimental Evidence

2.1 Method

We tested the grammatical judgments of the participants in a forced-choice, online violation detection task where the words were presented one at a time on the screen (RSVP), and the participants had to press a button to indicate whether a given sentence was grammatically correct or not.

Participants Fifty-four native speakers of English took part in our experiment, which was advertised on social media and mailing lists. The procedure and the consent were approved by the local ethical committee (Université Paris-Saclay, CER-Paris-Saclay-2019-063). We used filler trials (Table 1) to avoid potential confounding strategies from the participants, such as actively ignoring the middle of the sentences (Pearlmutter et al., 1999). Any participant whose answer to fillers was not significantly different from chance (binomial test, null hypothesis $p_0 = .5$) was rejected. We also rejected participants whose success rate on the main task was below 70%. Overall, we rejected 20 participants, and reported analyses from 34 participants (all analyses reported are consistent with corresponding analyses performed with the full dataset, reported in appendix B. For example, one can compare Table 2 and Table 5, and Figure 2 and Figure 5).

We also tested two transformer models (Wolf et al., 2020): a replication of the GPT-3 language model (Brown et al., 2020) made available by EleutherAI[1] (Black et al., 2021) and a Text-To-Text Transformer (T5)[2] (Raffel et al., 2019) fine-tuned on a grammatical error correction benchmark (Napoles et al., 2017). To evaluate the GPT-3 model, we input it with the sentence up to (excluding) the target verb and compare the probabilities associated with the grammatical and ungrammatical tokens (e.g., "climb" vs "climbs" for sentences in Table 1). Thus, for this model, we get a comparative performance per condition but cannot evaluate performances between grammatical and ungrammatical sentences. On the other hand, we compare humans directly with the grammaticality judgment of T5.

Experimental Procedure Participants were given a description of a violation-detection task (see Appendix A for the exact wording) includ-

[1] https://huggingface.co/EleutherAI/gpt-neo-1.3B

[2] https://huggingface.co/vennify/t5-base-grammar-correction

ing which button to press and three examples of grammatical and ungrammatical sentences. We presented sentences to participants, with a fixation cross between words, in white on black background, using a presentation time of 200ms and an SOA of 366ms. Participants could answer at any point by pressing the keyboard to indicate their judgment of grammaticality (left and right arrow keys randomized across subjects). Participants received auditory and visual feedback with each trial: green/red fixation and upward/downward tune (correct/incorect). At the end of the experiment, participants answered questions about the experiment (difficulty, weirdness, strategies), we provided them with an overall score and invited them to share the experiment on social media.

Stimuli We generated sentences from a fixed lexicon balanced for low-level features, yielding many sentences. Then we filtered them based on the perplexity of GPT-3, keeping only sentences that had an overall perplexity between the median and median $+2std$ — in order to keep sentences of low and consistent perplexity. We sampled 5 sentences for each condition (baseline, distal congruent, distal incongruent, proximal congruent, proximal incongruent), subject number (singular, plural), and grammaticality, as well as 32 filler sentences. No two sentences shared 5 words or more. The same sentences were presented to participants in randomized order.

2.2 Results

We call *incongruent* the trials in which the numbers of N_1 and N_2 disagree (Figure 1) — which is independent from *grammaticality*. Figure 2 shows the main effect of the presence of an attractor, and how its distance to the target (distal or proximal) modulates the performances of humans and NLMs.

In all cases, the baseline elicits fewer errors (ER) and faster reaction times (RT) compared to the attractor conditions. Errors occur more often in ungrammatical sentences than in grammatical ones, irrespective of the condition: this phenomenon is called *grammatical illusions* (Phillips et al., 2011) and indicates that participants accept ungrammatical sentences more often than they reject grammatical ones. We also replicate *grammatical asymmetry* (Wagers et al., 2009): incongruent trials lead to higher ER in the ungrammatical sentences.

To investigate the attractor effects, we analyzed the main factors of our design. The main effect

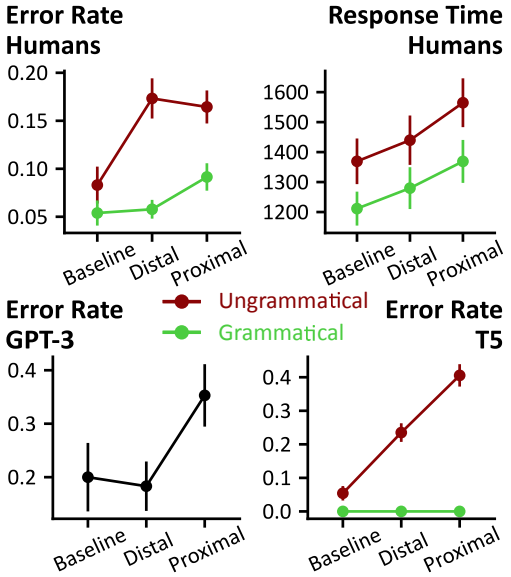

Figure 2: Performances of humans and NLMs: colors indicate grammaticality, error bars indicate (all figures) SEM over participants (humans) or sentences (NLMs).

of *Violation* refers to the dependency that controls the grammatical configuration of the sentence (Figure 1) and we use this effect as a proxy into structural processing (Rizzi, 2004). The *Congruency* effect refers to a dependency realized between two non-structurally related words and is used as a proxy for linear processing.

Figure 3 shows the main factors and their interaction. We report the corresponding ANOVA for the human participants in Table 2. We replicate the markedness phenomenon (Bock and Miller, 1991; Wagers et al., 2009): attraction effects surface only with plural attractors (see Table 4 and Figure 4 for analyses with singular attractors). Results shown in Figure 3 and Table 4 correspond to sentences with plural attractors.

In the human data, the main effect of *Violation* is significant across all conditions and dependent variables. Nevertheless, for the ER, the η^2_G value is larger by an order of magnitude in the distal attractor condition: this illustrates that participants make more mistakes in judging the grammaticality of sentences in which the attractor is far from the verb, especially when the sentence is ungrammatical. The distance of the attractor does not affect the magnitude of the effect in the case of RTs.

To elucidate this facilitatory effect, we focus on *Congruency*: its effect is only significant in the proximal condition, which illustrates that participants make more errors in incongruent sentences compared to the congruent ones (figure 3).

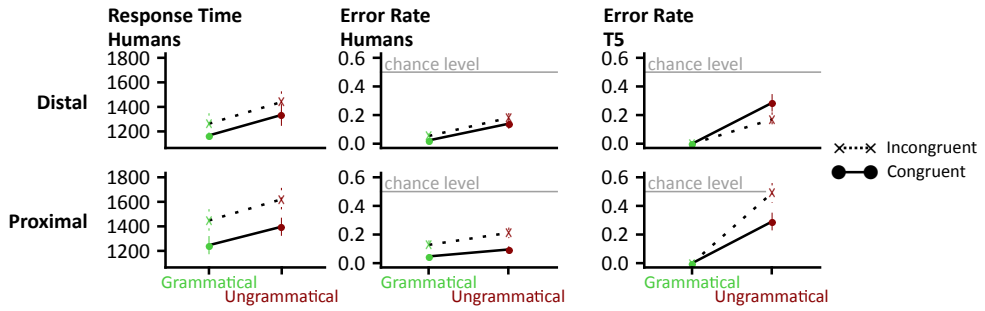

Figure 3: Effect of grammaticality, congruency, and distance of the attractor on our dependent variables.

Effect	$F_{1,33}$	p-value	η^2_G
Response Time; Distal attractor			
congruency	7.04	.012	.011
violation	11.64	.002	.031
interaction	0.01	.915	$< .001$
Response Time; Proximal attractor			
congruency	26.90	$< .001$	.046
violation	11.81	.002	.027
interaction	0.10	.752	$< .001$
Error Rate; Distal attractor			
congruency	2.29	.140	.013
violation	18.01	$< .001$	.135
interaction	0.04	.846	$< .001$
Error Rate; Proximal attractor			
congruency	15.79	$< .001$	.085
violation	5.33	.027	.040
interaction	0.57	.454	.003

Table 2: One-way between-subjects ANOVAs were conducted to compare the effect of congruency, violation, and their interaction on RT and ER for both proximal and distal attractors. Highlighted rows indicate significance at the $p < .05$ level. The last column provides η^2_G values, an estimator of the variance explained by the ANOVA, similar to r^2 for linear models.

NLMs displayed super-human performances in evaluating grammatical sentences. T5 is sensitive to the mere presence of an attractor, whose distance strongly modulates performance. GPT-3 is not sensitive to the mere presence of an attractor, but elicits a significant distance effect. Both models, but not humans, have near-chance performances in the case of a proximal attractor. These results might have a two-fold interpretation. The presence of *grammatical illusions* in both humans and models might be informative on the role of training in linguistic performance: in real sentences, grammatical sentences vastly outnumber ungrammatical ones. We might therefore be describing a similar training bias between humans and NLMs. There is a common response profile between the NLMs, which demonstrates the sensitivity of transformer models to statistical regularities. This allows us to zoom in on T5, which is more directly comparable to humans, and investigate how the *Congruency* factor modulates the ER.

The reduction in the ER in the human data was traced to a facilitatory effect of congruency: congruent trials led to fewer errors in the proximal compared to the distal attractor. On the contrary, in T5, the incongruent trials yielded chance-level performances, but there was no distance effect in the congruent trials. This indicates that congruent sentences help humans in performing the task, but not NLMs. Additionally, incongruent sentences have a detrimental effect in the models, but no such effect is observed in humans. These observations point to a common sensitivity to the attractor-verb distance in both systems, but a fundamental difference as to the outcome of this sensitivity.

3 Discussion

In this study, we revisited the classical attraction effect in subject-verb number agreement in humans and neural networks and sought to assess the influence of the attractor-verb distance.

Our results draw a picture of a shared distance sensitivity between humans and NLMs, but also a fundamental difference in the weight of this sensitivity. We observed that the artificial system operates on the basis of word-level statistics, and is thus driven to chance-level performance in the presence of a deviant bigram (Figure 3-bottom right-Incongruent trials). On the contrary, the incongruency of the sentence leads to comparable error rates irrespective of the attractor distance in humans (Figure 3-central column). The significant effect of congruency observed in the human data is due to a facilitatory effect of congruency in judging grammaticality, and not an inhibitory effect of incongruency, unlike with the neural networks. This effect is mostly evident in the agrammatical sentences and can be decomposed as follows.

Consider the following two sentences:

(a) [*Congruent*]
The writers who like the happy young **editors cries**.

(b) [*Incongruent*]
The writers who like the happy young **editor cries**.

Participants and $T5$ made fewer errors in (a) compared to (b). Participants also made fewer errors in (a) when the attractor was further away. We thus observe a facilitatory effect. Notice that here we have the realization of the locally implausible bigram **"editors cries"**. One possible interpretation might be that the participants were lured into judging this sentence as invalid, based on the presence of this bigram. A facilitatory effect is hence observed because the sentence was indeed agrammatical.

The error rate for the incongruent cases like (b) remained unchanged for the humans, with respect to the distal attractor (see Figure 3-central column). Importantly, though, in this case, where a plausible bigram is found (**"editor cries"**), the NLM reached chance-level performance. Given the local agreement of the attractor with the verb, the NLM was lured into judging the sentence as grammatical, when the sentence was wrong, whereas humans were able to mostly overcome the attraction effect.

Thus, our results suggest that operations at the n-gram level might play a key role in explaining the observed phenomena, in both humans and NLMs.

Nevertheless, alternative explanations can be described. In many behavioral studies, significant effects of congruency have been reported in non-intervening attractor structures such as Object Relative Clauses (e.g.: The player that the [teacher encourage/s] climbs) (Wagers et al., 2009). The n-gram mechanism cannot explain attraction phenomena in this setup, something that the dominant model of attraction (cue-retrieval model, (Wagers et al., 2009)) can do. In this model, a memory mechanism is enabled upon a cue, that retrieves the number of the subject from the memory system and the errors can be attributed to retrieval interference. Under this interpretation, the memory representation attractors fade with distance and therefore the distal attractor does not compete for retrieval. In contrast, when the attractor is in the vicinity of the verb, similarity-based retrieval interference can occur, and thus, attraction effects can only be realized in that condition (McElree et al., 2003).

However, it is important to note that none of the dominant models of grammatical agreement (cue-based retrieval model & feature percolation) are complete, as both are conclusive in ungrammatical conditions, but not always for the grammatical ones. Thus, a need of model revision seems necessary. In this study, we tried to point to the calculation of transition probabilities as a candidate factor for model development.

3.1 Limitations

There are a number of limitations that narrows the scope of our results. First, it is difficult to draw general conclusions from a single experiment on grammatical number agreement. Second, although all conditions are balanced, the stimuli we used are not devoid of semantic content which might induce some biases. These results will need to be confirmed in future experiments using different tasks and stimuli, for example using morphosyntactically marked stimuli but devoid of semantics, so-called "jabberwocky" sentences (Hahne and Jescheniak, 2001; Desbordes et al., 2023). This subject represents a significant area for our ongoing research and exploration.

Third, we compared human participants to language models under the assumption that NLMs express sensitivity to probabilistic relationships at the word level, and thus a comparison under the same conditions might shed light on the processing of language in the human brain. We are fully aware that this comparison is indirect, and that an LSTM architecture might have been more appropriate for such a comparison. Nevertheless, the literature has pointed to differences between grammatical and ungrammatical conditions, effects that we replicated in this study, and we therefore sought for a model that would allow us for a direct comparison for each condition.

3.2 Conclusion

Our results show that, in humans and NLMs, language processing is affected by the attractor-verb distance. We additionally hypothesize that this is due to the calculation of transition probabilities at the word level, which can either run contrary or reinforce the overall structure-based processing.

Humans are less affected by this local interference, suggesting that language models process language in ways that are still fundamentally different from humans, even though they superficially coincide, e.g. in grammatical cases in our data.

References

Sophie L Arana, Jan-Mathijs Schoffelen, Tom Mitchell, and Peter Hagoort. 2021. Mvpa does not reveal neural representations of hierarchical linguistic structure in meg. *bioRxiv*.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Kathryn Bock and Carol A Miller. 1991. Broken agreement. *Cognitive psychology*, 23(1):45–93.

Jonathan R. Brennan, Edward P. Stabler, Sarah E. Van Wagenen, Wen-Ming Luh, and John T. Hale. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. 157-158:81–94.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Noam Chomsky. 1957. *Syntactic structures*. Mouton publishers. Accepted: 2019-04-10T06:25:21Z Journal Abbreviation: Janua linguarum.

Cas W Coopmans, Helen De Hoop, Karthikeya Kaushik, Peter Hagoort, and Andrea E Martin. 2021. Structure-(in) dependent interpretation of phrases in humans and lstms. In *The Society for Computation in Linguistics (SCiL 2021)*, pages 459–463.

Stanislas Dehaene, Florent Meyniel, Catherine Wacongne, Liping Wang, and Christophe Pallier. 2015. The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. *Neuron*, 88(1):2–19.

Théo Desbordes, Yair Lakretz, Valérie Chanoine, Maxime Oquab, Jean-Michel Badier, Agnès Trébuchon, Romain Carron, Christian-G. Bénar, Stanislas Dehaene, and Jean-Rémi King. 2023. Dimensionality and ramping: Signatures of sentence integration in the dynamics of brains and deep language models. *Journal of Neuroscience*. Publisher: Society for Neuroscience Section: Research Articles.

Nai Ding, Lucia Melloni, Hang Zhang, Xing Tian, and David Poeppel. 2015. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, 19(1):158–164.

Victoria Fossum and Roger Levy. 2012. Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 61–69, Montréal, Canada. Association for Computational Linguistics.

Julie Franck, Gabriella Vigliocco, and Janet Nicol. 2002. Subject-verb agreement errors in french and english: The role of syntactic hierarchy. *Language and cognitive processes*, 17(4):371–404.

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

Karl Friston, Rosalyn J. Moran, Yukie Nagai, Tadahiro Taniguchi, Hiroaki Gomi, and Josh Tenenbaum. 2021. World model learning and inference. 144:573–590.

A. Hahne and J. D. Jescheniak. 2001. What's left if the Jabberwock gets the semantics? An ERP investigation into semantic and syntactic processes during auditory sentence comprehension. *Brain Research. Cognitive Brain Research*, 11(2):199–212.

Christopher Hammerly, Adrian Staub, and Brian Dillon. 2019. The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive psychology*, 110:70–104.

Todd R Haskell and Maryellen C MacDonald. 2005. Constituent structure and linear order in language production: evidence from subject-verb agreement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5):891.

Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. 2002. The faculty of language: what is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579.

Jaap Jumelet, Willem Zuidema, and Dieuwke Hupkes. 2019. Analysing neural language models: Contextual decomposition reveals default reasoning in number and gender assignment. *arXiv preprint arXiv:1909.08975*.

Yair Lakretz, Théo Desbordes, Jean-Rémi King, Benoît Crabbé, Maxime Oquab, and Stanislas Dehaene. 2021. Can rnns learn recursive nested subject-verb agreements?

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.

R Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2021. How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *arXiv preprint arXiv:2111.09509*.

Brian McElree, Stephani Foraker, and Lisbeth Dyer. 2003. Memory structures that subserve sentence comprehension. *Journal of memory and language*, 48(1):67–91.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40.

Nicola Molinaro, Horacio A Barber, and Manuel Carreiras. 2011. Grammatical agreement processing in reading: Erp findings and future directions. *cortex*, 47(8):908–930.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. Jfleg: A fluency corpus and benchmark for grammatical error correction.

Matthew J Nelson, Imen El Karoui, Kristof Giber, Xiaofang Yang, Laurent Cohen, Hilda Koopman, Sydney S Cash, Lionel Naccache, John T Hale, Christophe Pallier, et al. 2017. Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 114(18):E3669–E3678.

Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. 2011. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527.

Neal J Pearlmutter, Susan M Garnsey, and Kathryn Bock. 1999. Agreement processes in sentence comprehension. *Journal of Memory and language*, 41(3):427–456.

Colin Phillips, Matthew W Wagers, and Ellen F Lau. 2011. Grammatical illusions and selective fallibility in real-time language comprehension. *Experiments at the Interfaces*, 37:147–180.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Luigi Rizzi. 2004. On the cartography of syntactic structures. *The Structure of CP and IP*, pages 3–15.

Rushen Shi, Camille Legrand, and Anna Brandenberger. 2020. Toddlers track hierarchical structure dependence. *Language Acquisition*, 27(4):397–409.

Gabriella Vigliocco and Janet Nicol. 1998. Separating hierarchical relations and word order in language production: Is proximity concord syntactic or linear? *Cognition*, 68(1):B13–B29.

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.

Jana Willer Gold, Boban Arsenijević, Mia Batinić, Michael Becker, Nermina Čordalija, Marijana Kresić, Nedžad Leko, Franc Lanko Marušič, Tanja Milićev, Nataša Milićević, et al. 2017. When linearity prevails over hierarchy in syntax. pnas.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Appendices

A Questionnaire, Instructions, and Stimuli

Instructions

The exact instructions given to the participants are provided below. They consisted of three separate pages, participants could go back and forth between pages freely.

Remember that the key/response binding was randomized across subjects, so page 3 provided below only applied to half of our participants, where the other half had a corresponding, flipped association of key and responses.

Page 1

- This experiment is about sentence processing
- You will read sentences on the screen, with the words presented one after the other, at the center of the screen
- Some of these sentences will contain mistakes
- Your task is to find these mistakes

Page 2

Here are a few examples to show what we mean by correct and incorrect. Remember that the sentence will not be presented as a whole, but rather one word after another.
Incorrect examples:

- The boy drink water while listening to music
- The farmer near the two pilot detests boxing
- The athletes that dislike the happy proud banker sings

Correct examples:

- The boy drinks water while listening to music
- The farmer near the two pilots detests boxing
- The athletes that dislike the happy proud banker sing

Some sentences might be a bit weird, like in example 3, but you should always be able to perform the task if you remain focused.

Page 3

You have to look at the cross at the center of the screen, which is always present when there is no word to read. Make sure the luminosity of your screen is high enough for you to read. Then you will read sentences one word after the other and you have to do the following:

- As soon as you think a given sentence is INCORRECT, please press the -> right arrow key on your keyboard
- When the sentence ends, if you think it is CORRECT, please press the <- left arrow key on your keyboard
- You have to answer every time, even when you're not sure or you feel you don't know. Only after you answer, the following sentence will start. Answer the best you can!

After each answer you will receive feedback: the central cross will turn green if you answered correctly, and red otherwise. If you can, please turn your computer audio on: that way, you will receive feedback with sounds for each trial.
This is the last instruction page. You can go back to the other pages, but when you move forward the experiment ask you to go fullscreen. Then the experiment will start with 5 training examples so that you understand the task.

Material

All stimuli are provided below, first the grammatical ones (1-50), then the ungrammatical ones (51-100):

1. The lawyers who avoid the kind gentle judge lie.
2. The athlete who hates the farmers the least sings.
3. The judges that fear the proud charming man sing.
4. The builder who dislikes the proud gentle farmer cheats.
5. The plumbers that run happily although rather quickly lie.
6. The painter who loves the young lazy farmers cheats.
7. The waiter who avoids the judge the least cooks.
8. The waiters that love the chefs the least pray.
9. The tailor who avoids the farmer the least prays.
10. The waiters that walk happily albeit pretty quickly pray.
11. The judges that avoid the tailors the most swim.
12. The women who love the happy clever chefs sing.
13. The lawyer that runs carefully yet fairly quickly swims.
14. The athlete that loves the vet the most lies.
15. The waiters who walk carefully yet pretty quickly lie.
16. The teacher who fears the lawyers the most cheats.
17. The teachers who fear the plumber the most climb.
18. The waiters who avoid the clumsy clever plumber cheat.
19. The waiter who dislikes the proud nice woman swims.
20. The chefs who avoid the waiters the least climb.
21. The painters who fear the funny nice women pray.
22. The vets that like the farmer the most smoke.
23. The doctor that runs happily albeit rather carefully cheats.
24. The teachers that dislike the tailors the least cheat.
25. The lawyers that love the waiter the least climb.
26. The plumber who fears the lawyer the most climbs.
27. The painters that love the careless proud judges smoke.
28. The farmers that run happily though fairly quickly sing.
29. The waiters that walk carefully although rather quickly cheat.
30. The man who laughs happily though pretty quickly lies.

31. The plumber that rides happily although pretty quickly swims.
32. The actor that dislikes the lawyer the most prays.
33. The chef who dislikes the authors the least cheats.
34. The vet that likes the proud helpful painters cheats.
35. The farmer who fears the clever lazy tailors cheats.
36. The painter who dislikes the nice careless teacher cheats.
37. The painters that dislike the careless helpful tailors climb.
38. The waiters who avoid the men the most sing.
39. The actors that hate the painter the most cook.
40. The teacher who likes the bakers the most sings.
41. The actor who dislikes the chefs the most swims.
42. The lawyers that fear the funny kind athlete pray.
43. The bakers that fear the man the least pray.
44. The actors who fear the nice proud men cook.
45. The man who laughs carefully yet rather quickly smokes.
46. The athlete who hates the proud funny woman prays.
47. The tailor that loves the clever clumsy bakers cheats.
48. The man who avoids the clumsy helpful chef sings.
49. The vets who dislike the young nice man cheat.
50. The man who fears the lazy nice authors lies.
51. The baker who likes the clever happy plumber swims.
52. The baker that hates the lazy gentle man cooks.
53. The waiter who likes the actors the most smokes.
54. The tailors who like the nice cool men swim.
55. The women who fear the lazy clumsy plumber sing.
56. The baker who dislikes the judge the least cooks.
57. The woman that fears the baker the most cheats.
58. The plumber who talks happily yet rather quickly swims.
59. The lawyer that likes the farmers the most swims.
60. The bakers that love the careless clever chef pray.
61. The man who walks carefully although fairly quickly cheats.
62. The man that runs carefully though rather quickly lies.
63. The bakers who love the actor the least cheat.
64. The tailor that dislikes the cool lazy baker prays.
65. The plumbers who love the kind charming doctor cheat.
66. The pilots that hate the waiter the most cook.
67. The baker who dislikes the kind helpful plumbers lies.
68. The farmer who fears the doctors the least swims.
69. The doctor who hates the actor the least cheats.
70. The waiter who hates the cool clumsy man swims.
71. The farmer who likes the builders the least prays.
72. The waiter who dislikes the painter the least prays.
73. The woman who avoids the gentle lazy waiters prays.
74. The author that hates the waiters the least smokes.
75. The doctor that hates the careless young teacher lies.
76. The chef that dislikes the proud clumsy tailors sings.
77. The athletes who love the judges the least sing.
78. The bakers that love the lovely helpful women cook.
79. The tailors that drive happily although fairly quickly climb.
80. The tailors who run carefully yet fairly happily cheat.
81. The chefs who love the proud cool bakers cook.
82. The bakers who fear the woman the most climb.
83. The teacher who dislikes the helpful charming builders cheats.
84. The chefs who run carefully albeit rather quickly cook.
85. The painter who loves the helpful friendly judges prays.
86. The waiters that run carefully though pretty quickly cook.
87. The teachers that avoid the waiters the most pray.
88. The waiters who drive happily yet fairly quickly swim.
89. The painter that avoids the waiter the most prays.
90. The tailors that love the charming young authors smoke.
91. The lawyers that dislike the proud young farmer cheat.
92. The vets that love the woman the least cook.
93. The doctors that like the bakers the least pray.
94. The builder that drives happily though rather quickly cheats.
95. The plumbers who dislike the careless clever bakers sing.
96. The athletes who hate the bakers the most swim.
97. The doctors that like the chef the most cook.
98. The bakers who fear the friendly nice man swim.
99. The men who dislike the pilots the least pray.
100. The plumber that laughs carefully yet pretty quickly prays.

B Additional Results

Table 3: ANOVAs for two dependent variables in humans (response time and error rates) and for error rates in T5, testing whether there was a significant effect of violation in three possible conditions: Baseline, (no attractor), and Distal and Proximal attractors.

condition	Statistic	p-value	η^2_G
Response Time (humans)			
Baseline	$F_{1,33} = 1.81$	.187	.023
Distal	$F_{1,33} = 27.86$	$< .001$	.274
Proximal	$F_{1,33} = 13.22$	$< .001$	.139
Error Rate (humans)			
Baseline	$F_{1,33} = 10.91$	.002	.040
Distal	$F_{1,33} = 22.76$	$< .001$	.032
Proximal	$F_{1,33} = 28.60$	$< .001$	.047
Error Rate (T5)			
Baseline	$F_{1,110} = 6.29$	.014	.054
Distal	$F_{1,233} = 71.59$	$< .001$	.235
Proximal	$F_{1,216} = 147.35$	$< .001$	.406

Table 4: Corresponding table to Table 1 with the filter operation when the attractor is singular.

Effect	$F_{1,33}$	p-value	η^2_G
Response Time; Distal attractor			
congruency	1.12	.297	.001
violation	14.02	$< .001$	.021
interaction	0.15	.704	$< .001$
Response Time; Proximal attractor			
congruency	0.00	.961	$< .001$
violation	13.66	$< .001$	.045
interaction	0.32	.578	$< .001$
Error Rate; Distal attractor			
congruency	0.03	.867	$< .001$
violation	14.52	$< .001$	.091
interaction	0.88	.356	.003
Error Rate; Proximal attractor			
congruency	0.00	$> .999$	$< .001$
violation	8.17	.007	.053
interaction	0.38	.539	.002

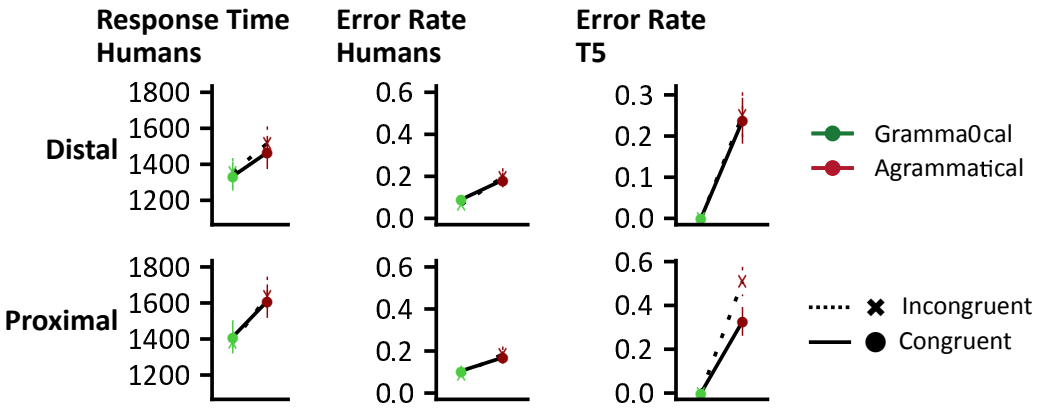

Figure 4: Corresponding figure to 3 filtering for singular attractors: observe the overall collapse of the effect of congruency.

Table 5: Table 2 with full dataset of N=54 participants included

Effect	$F_{1,53}$	p-value	η^2_G
Response Time; Distal attractor			
congruency	0.14	.714	$< .001$
violation	11.77	.001	.024
interaction	0.08	.773	$< .001$
Response Time; Proximal attractor			
congruency	0.16	.691	$< .001$
violation	17.98	$< .001$	.038
interaction	0.97	.329	.002
Error Rate; Distal attractor			
congruency	0.01	.918	$< .001$
violation	20.70	$< .001$	.062
interaction	0.09	.768	$< .001$
Error Rate; Proximal attractor			
congruency	0.75	.391	.002
violation	18.21	$< .001$	.064
interaction	0.27	.607	$< .001$

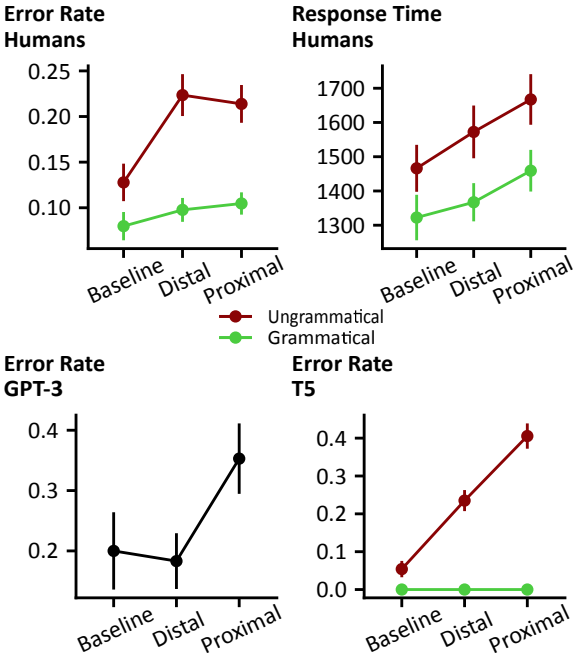

Figure 5: Performances of humans and NLMs, using all participants: colors indicate grammaticality, error bars indicate (all figures) SEM over participants (humans) or sentences (NLMs). Results are comparable to the ones after rejection of low-performing participants