Neural network modelling on Korean monolingual children's comprehension of suffixal passive construction in Korean

Seongmin Mun (stat34@ajou.ac.kr)
Humanities Research Institute, Ajou University
206 World cup-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, South Korea

Gyu-Ho Shin (ghshin@uic.edu)
Department of Linguistics, University of Illinois at Chicago
601 South Morgan Street, Chicago, IL 60607 USA

Abstract

This study explores a GPT-2 architecture's capacity to capture monolingual children's comprehension behaviour in Korean, a language underexplored in this context. We examine its performance in processing a suffixal passive construction involving verbal morphology and the interpretive procedures driven by that morphology. Through model fine-tuning via patching and hyperparameter variations, we assess their classification accuracy on test items used in Shin (2022a). Results show discrepancies in simulating children's response patterns, highlighting the limitations of neural networks in capturing child language features. This prompts further investigation into computational models' capacity to elucidate developmental trajectories of child language that have been unveiled through corpus-based or experimental research.

Keywords: Neural network; Child comprehension; Passive construction; Korean

Introduction

One notable trend in language sciences is to apply computational methods and techniques to pursue linguistic inquiries. This line of research has explored computational models' capacity to simulate human language behaviour (Hawkins et al., 2020; Marvin & Linzen, 2019; Warstadt et al., 2019), together with performance-wise variations across algorithms (Hu et al., 2020), thereby gaining momentum in addressing how learning occurs in the human mind without presuming innate knowledge about grammar (Contreras Kallens et al., 2023; O'Grady & Lee, 2023; Perfors et al., 2011; Warstadt & Bowman, 2020). An emerging—but less active—strand of research applies computational methods and techniques to reveal developmental trajectories of linguistic knowledge (Alishahi & Stevenson, 2008; Ambridge et al., 2020; Bannard et al., 2009; Chang, 2009; Sagae, 2021; You et al., 2021). Despite its significance, the current research practice bears three grave limitations. First, the field is skewed heavily towards a limited range of languages (and especially English) and language-usage contexts (e.g., adult language). In particular, based on the predominance of English-oriented Large Language Models (LLMs), the intensification of this research bias is being accelerated. This restricts the generalisability of findings from previous studies to lesser-studied languages and registers. Second, while the vast majority of work on this topic seeks to propose new models or improve currently available models, researchers pay relatively little attention to whether and how the implications of computational simulations are compatible with those of other types of measurement, such as behavioural experiments and corpus findings revealing fundamental architectures of human language behaviour. We are aware of few studies informative in this regard (Ambridge et al., 2020; Oh et al., 2022; Xu et al., 2023). Third, researchers' access to computing resources in academia is limited. Researchers in academia often confront costly access to cutting-edge algorithms and pre-trained models, as well as weak computing power. Together, these limitations pose a serious threat to diversity, equity, and inclusion in research (cf. Bender et al., 2021).

This study aims to alleviate these concerns by exploring how neural networks capture children's comprehension behaviour. We adopt GPT-2 (Radford et al., 2019) for our inquiry. GPT utilises attention for effective computation by enhancing each part of the input sequence in consideration of various information about the whole sequence (e.g., segment position) to better identify the most relevant parts of that sequence (Vaswani et al., 2017). Because this algorithm targets a general-purpose learner whose learning trajectories are not subject to particular tasks, model training does not stand on the specifics of data or tasks (Radford et al., 2019); it can also perform new tasks with a relatively small number of examples. Despite the continuous development of the GPT-*n* architecture, GPT-2 is often employed to conduct simulations on language behaviour (Goldstein et al., 2022; Hosseini et al., 2022), yielding successful modelling on language tasks.

Acquisition of suffixal passive in Korean

Korean is an agglutinative, SOV language with overt case-marking via dedicated particles and active use of verbal morphology to indicate grammatical information. This language is understudied in the field and is computationally challenging due to its language-specific properties. Two major clausal constructions deliver transitivity in Korean: active transitive and passive. The canonical active transitive pattern in Korean, when fully marked as in (1a), occurs with a nominative-marked agent, followed by an accusative-marked theme; a verb carries no dedicated active morphology. Korean allows scrambling of sentential components as in (1b) if that reordering (agent–theme → theme–agent) preserves the basic propositional meaning. In addition, omission of sentential components is permitted as in (1c-d) if event participants are clearly identified with no ambiguity arising within the context

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

(Sohn, 1999). Of the three types of passive construction, the suffixal passive (which is the most frequent type that children encounter; Shin & Mun, 2023a) consists of two arguments, a nominative-marked theme and a dative-marked agent occupying the subject and oblique positions, respectively; a verb carries dedicated passive morphology as one of the four allomorphic variants of suffixes *-i-, -hi-, -li-,* or *-ki-*. While the canonical pattern follows the theme–agent ordering as in (2a), it can be scrambled, yielding the agent–theme ordering as in (2b) with the propositional meaning intact.

(1) Active transitive: 'Mina hugged Pola.'
  a. Canonical
    Mina-ka Pola-lul an-ass-ta.
    Mina-NOM Pola-ACC hug-PST-SE[1]
  b. Scrambled
    Pola-lul Mina-ka an-ass-ta.
    Pola-ACC Mina-NOM hug-PST-SE
  c. Omission (case marker)
    Mina-ka Pola-~~lul~~ an-ass-ta.
    Mina-NOM Pola-~~ACC~~ hug-PST-SE
  d. Omission (argument & case marker)
    ~~Mina-ka~~ Pola-lul an-ass-ta.
    ~~Mina-NOM~~ Pola-ACC hug-PST-SE

(2) Suffixal passive: 'Pola was hugged by Mina.'
  a. Canonical
    Pola-ka Mina-hanthey an-ki-ess-ta.
    Pola-NOM Mina-DAT hug-PSV-PST-SE
  b. Scrambled
    Mina-hanthey Pola-ka an-ki-ess-ta.
    Mina-DAT Pola-NOM hug-PSV-PST-SE

Passive morphology serves as a key disambiguation point to identify the structural properties of the suffixal passive, forcing a comprehender to revise the initial analysis prior to that morphology. In Korean, a nominative-marked [+human] argument is likely to be interpreted as an agent, and a dative-marked [+human] argument is likely to be interpreted as a recipient; these interpretations are supported by strong mapping between thematic roles and case markers attested in language use (Kim & Choi, 2004; Sohn, 1999). Therefore, a plausible way of analysing (2) prior to the verb is that *Pola* acts on/for *Mina*. However, this initial analysis is incongruent with the passive-voice information conveyed by verbal morphology. Thus, upon encountering the verb at the sentence-final position, a comprehender must revise the initial interpretation by recalibrating the arguments' thematic roles, mapping a theme role onto the nominative-marked entity and an agent role onto the dative-marked entity. This revision process is demanding (Kendeou et al., 2013; Rapp & Kendeou, 2007; Trueswell et al., 1999), adding difficulty in children's comprehension of this construction (Shin, 2022a; Shin & Deen, 2023; Kim et al., 2017).

Shin (2022a), the baseline of this study, explored Korean monolingual children's comprehension behaviour involving the suffixal passive construction through four picture-selection experiments combined with a novel methodology that systematically omitted or obscured portions of test sentences using acoustic sounds (e.g., cough, chewing). In each experiment, a pair of two pictures was presented involving the same action but reversed thematic roles (e.g., a dog kicking a cat; a cat kicking a dog), and a sentence indicating one of the two pictures (e.g., *kangaci-ka koyangi-hanthey cha-i-eyo*. dog-NOM cat-DAT kick-PSV-SE 'The dog is kicked by the cat.') was presented twice orally; participants (3-4yrs; 5-6yrs; adults) were asked to choose a picture that matched the sentence.

The four experiments generated three major findings on the children's comprehension of the suffixal passive (Table 1). First, given the competition between passive-voice knowledge (induced by verbal morphology) and active-voice knowledge (which is frequent in use and well-entrenched in children's minds), utilising passive-voice knowledge was subject to age (as a proxy for language-usage experience). Second, the 5-6yrs were able to deploy passive-voice knowledge, but the degree to which they employed that knowledge was inversely proportional to the computational complexity of a sentence (e.g., number of arguments, type of case markers present/absent). Third, the 3-4yrs did not fully respect active-like interpretation when comprehending the passive sentences. These findings indicate an emerging sensitivity to passive morphology and an increasing capacity to utilise passive-voice knowledge tied to that morphology with age, in conjunction with the interplay between voice-related knowledge involving a given stimulus. This suggests early emergence, but late mastery, of linguistic knowledge, the maturation of which necessitates a significant amount of language-usage experience.

Table 1. Summary of experimental results: Shin (2022a)

Exp	Condition	3-4yrs		5-6yrs		Adult	
		Mean	SD	Mean	SD	Mean	SD
1	$N_{NOM}N_{ACC}V_{act}$	0.844	0.36	0.942	0.24	1.000	0.00
	$N_{ACC}N_{NOM}V_{act}$	0.778	0.42	0.710	0.46	1.000	0.00
	$N_{NOM}N_{DAT}V_{psv}$	0.456	0.50	0.478	0.50	1.000	0.00
	$N_{DAT}N_{NOM}V_{psv}$	0.511	0.50	0.768	0.43	1.000	0.00
2	$N_{\sim\sim CASE}N_{\sim\sim CASE}V_{act}$	0.667	0.48	0.773	0.42	0.900	0.30
	$N_{\sim\sim CASE}N_{\sim\sim CASE}V_{psv}$	0.545	0.50	0.424	0.50	0.150	0.36
3	$N_{NOM}V_{act}$	0.944	0.23	0.971	0.17	0.933	0.25
	$N_{ACC}V_{act}$	0.922	0.27	0.971	0.17	1.000	0.00
	$N_{NOM}V_{psv}$	0.522	0.50	0.710	0.46	0.967	0.18
	$N_{DAT}V_{psv}$	0.533	0.50	0.841	0.37	0.950	0.22
4	$N_{\sim\sim CASE}V_{act}$	0.426	0.50	0.604	0.50	0.667	0.48
	$N_{\sim\sim CASE}V_{psv}$	0.593	0.50	0.333	0.48	0.100	0.30

Note. 'Mean' denotes the average accuracy (in Exps 1 and 3) or the average rate of agent-first response (Exps 2 and 4).

[1] Abbreviations: ACC = accusative case marker; DAT = dative marker; NOM = nominative case marker; PSV = passive suffix; PST = past tense marker; SE = sentence ender; Strikethrough in grey = obscured; V = verb.

Methods[2]

Great interest lies in the ability of neural network models to recognise passive morphology and execute the necessary revision process for correctly interpreting suffixal passive sentences. We developed GPT-2 models by (i) fine-tuning via patching (i.e., pre-trained model + caregiver input) and (ii) adjusting hyperparameters, and assessed their classification performance on test stimuli identical to those used in Shin (2022a). Unlike other studies incorporating additional variables such as thematic role variables (Chang, 2002) and a separate layer encoding semantic information (Alishahi & Stevenson, 2008), our models solely engaged with formal features (i.e., raw text) during training and classification. This study also extends Shin & Mun (2023b), exploring how hyperparameter variations influence model performance with respect to child language data.

Data Pre-processing

We used caregiver input in CHILDES (MacWhinney, 2000), adopting the same data used in Shin & Mun (2023b) considering the comparability of findings between Shin & Mun (2023b) and this study. The data were pre-processed by (i) correcting typos and spacing errors and (ii) excluding any sentence whose length was less than five characters or those consisting only of onomatopoeic and mimetic words. This treatment resulted in 69,498 sentences (285,350 eojeols[3]).

Model Training

Table 2 provides details on the models created in this study.

Table 2: Specification of computational models.

Python Package	*Transformers* (version 4.35)
Pre-trained model	*KoGPT2-base-v2* (Size: 51,200)
Tokenisation	Syllable-based; *Byte Pair Encoding*
Internal setting	Epoch: 10, Seed: 42, Epsilon: 0.00000001 Embedding/hidden dimension: 768 FFN inner hidden dimension: 3,072 Attention head #: 12, Parameter #: 125M Transformer layer #: 12
Hyperparameter variation	Learning rate: 0.001, 0.0001 Batch size: 16, 64 / Max. SeqLen: 64, 256

Neural networks typically require large-scale data for training to ensure their optimal operation (Edwards, 2015), but there is no pre-trained model exclusively constructed with caregiver input nor enough Korean caregiver-input data to create a pre-trained model. In addition, children encounter more than just caregiver input in real life; there are many types of exposure to language use with which children are surrounded. To cope with these issues, we employed a pre-trained model, which was open-access and representative at the moment of study, and patched the caregiver-input data to the pre-trained model when developing our models. The patching procedure increased the pre-trained model's size (51,200 to 67,052). We believe that adopting a pre-trained model in conjunction with the caregiver-input data can improve ecological validity for this type of modelling, but no research has scrutinised this point thoroughly, indicating the need for further attention. We also manipulated three hyperparameters to test whether and how hyperparameter variations influence model performance when handling child language data. Our choices were informed by previous studies (Budzianowski & Vulić, 2019; Dai et al., 2023; Oh & Schuler, 2022; de Vries & Nissim, 2021). These variations generated 8 sub-models.

For the binary classification of test items (*Agent-First; Theme-First*), our models were further fine-tuned on instances of all the constructional patterns expressing a transitive event—active transitive and suffixal passive, with scrambling and varying degrees of omission manifested—with labels indicating if the thematic-role ordering of these instances followed agent-first or theme-first. The instances were extracted from the caregiver-input data in CHILDES through an automatic search process developed by Shin (2022b); every sentence for each extraction was checked manually to confirm its accuracy. This treatment also aimed to ensure the compatibility between the simulation environments and the experimental settings of Shin (2022a), in which participants were presented transitive-event pictures prior to a stimulus to contextualise their interpretation of that stimulus. Furthermore, considering the zero occurrence of some patterns in the input, we adapted the Laplace smoothing technique (Agresti & Coull, 1998) by adding one fake instance to all the patterns.

As illustrated in Figure 3, each input sentence in the fine-tuning stage was transformed into two embedding types.

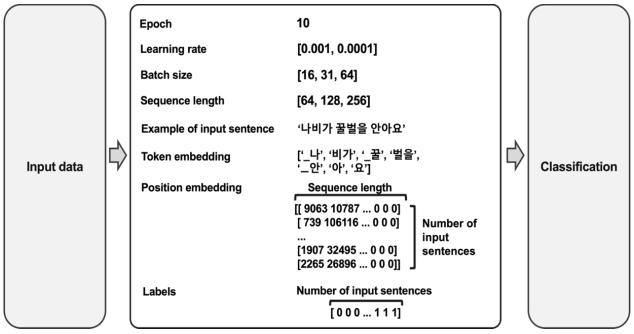

Figure 1: Model training (example sentence: *napi-ka kkwulpel-ul an-ayo* butterfly-NOM honeybee-ACC hug-SE 'The butterfly hugs the honeybee.').

For token embedding, the sentences were tokenised as syllable units. Originally, GPT-2 utilised a character for this task in the case of English. However, KoGPT-2 employs a syllable as a basic unit of tokenisation, likely in consideration of the language-specific properties of Korean. For position embedding, each token was converted into a numeric value indicating a unique index of the token with reference to the

[2] See this repository for the code and dataset.

[3] An eojeol refers to a unit with whitespace on both sides that serves as the minimal unit of sentential components. This roughly corresponds to a word in English.

vocabulary in the patched pre-trained model. The maximum dimension size of position embeddings was determined by the maximum sequence length set in the hyperparameter-setting stage. The initial values of epsilon and seed were automatically updated with the outcomes of each epoch. The training occurred from the initial model with the zero value of gradients to an optimal model with updated values through feedforward and backpropagation. After the training, the model evaluated the test stimuli, accumulating by-syllable information sequentially (by generating respective hidden layers) and then comparing the outcomes (1 = *Agent-First*; 0 = *Theme-First*) to the actual labels of these stimuli. We repeated the same learning process 30 times in each epoch and averaged the by-condition outcomes in assessing the models' classification performance to alleviate potential variations during the task.

Model Evaluation

For test items, we employed the same stimuli used in Shin (2022a). Each condition comprised six instances, with animals as agents and themes and actional verbs at the end (Table 3). Each trained model classified every test stimulus, evaluating whether the stimulus fell into *Agent-First* or *Theme-First*.

Table 3: Composition of test stimuli.

Condition	Example	Expected classification
$N_{NOM}N_{ACC}V_{act}$	cat-NOM dog-ACC kick	Agent-first
$N_{ACC}N_{NOM}V_{act}$	dog-ACC cat-NOM kick	Theme-first
$N_{NOM}N_{DAT}V_{psv}$	cat-NOM dog-DAT kick-PSV	Theme-first
$N_{DAT}N_{NOM}V_{psv}$	dog-DAT cat-NOM kick-PSV	Agent-first
$N_{CASE}N_{CASE}V_{act}$[a]	cat dog kick	Agent-first
$N_{CASE}N_{CASE}V_{psv}$[a]	cat dog kick-PSV	Theme-first
$N_{NOM}V_{act}$	cat-NOM kick	Agent-first
$N_{ACC}V_{act}$	dog-ACC kick	Theme-first
$N_{NOM}V_{psv}$	cat-NOM kick-PSV	Theme-first
$N_{DAT}V_{psv}$	dog-DAT kick-PSV	Agent-first
$N_{CASE}V_{act}$[a]	dog kick	Agent-first
$N_{CASE}V_{psv}$[a]	dog kick-PSV	Theme-first

We note that, while the stimuli in the case-less conditions in Shin (2022a) involved acoustic masking effects, the same stimuli in the simulations did not have such auditory effects. This was unavoidable considering this study's simulation setting, in which the models worked exclusively with the textual data. We concede that this difference may serve as one confounding factor for interpreting the results.

Results and Discussion

Case-marked Conditions

Two-argument conditions In $N_{NOM}N_{ACC}V_{act}$, all the models demonstrated high accuracy as the epoch progressed. However, in $N_{ACC}N_{NOM}V_{act}$, the models' accuracy rates were close to 0. This indicates that the models classified the test stimuli in this condition into *Agent-First* most of the time (which should have been *Theme-First*), resembling those of Shin and Mun (2023b). In $N_{NOM}N_{DAT}V_{psv}$, all the models' accuracy rates were close to 0. This indicates that the models classified the test stimuli in this condition into *Agent-First* most of the time (which should have been *Theme-First*). However, in $N_{DAT}N_{NOM}V_{psv}$, all the models demonstrated high accuracy as the epoch progressed.

One-argument conditions In both $N_{NOM}V_{act}$ and $N_{ACC}V_{act}$, all the models demonstrated high accuracy as the epoch progressed. In $N_{NOM}V_{psv}$, the models showed very low accuracy, regardless of hyperparameter type. This indicates that they classified the test stimuli in this condition into *Agent-First* most of the time (which should have been *Theme-First*). In $N_{DAT}V_{psv}$, except the models with the learning rate of 0.0001, all the models demonstrated high accuracy as the epoch progressed.

Case-less Conditions (coded as *Agent-First* = 1)

In both $N_{CASE}N_{CASE}V_{act}$ and $N_{CASE}N_{CASE}V_{psv}$, the models were at-chance or slightly above-chance. In $N_{CASE}V_{act}$, the models were at-chance or slightly below-chance, independently of hyperparameter types, which aligns with Shin & Mun (2023b) but not with Shin (2022a). In $N_{CASE}V_{psv}$, whereas the models with a learning rate of 0.001 showed at-chance performance, the models with a learning rate of 0.0001 showed below-chance performance, indicating that they classified the test stimuli in this condition as *Theme-First* most of the time.

Discussion

While the GPT-2 models' performance converged with the children's response patterns found in Shin (2022a) to some extent, the models did not faithfully simulate the children's comprehension behaviour pertaining to the suffixal passive, which yielded by-condition/hyperparameter asymmetries.

The results of this study are attributable to various factors. For instance, whereas Korean caregiver input joins the general characteristics of child-directed speech (Shin, 2022b; cf. Cameron-Faulkner et al., 2003; Stoll et al., 2009), it also manifests language-specific properties, such as scrambling and omission of sentential components. The models may thus have been sensitive to the specific word order and the type of case markers present in a stimulus during the classification task, particularly as shown in $N_{ACC}N_{NOM}V_{act}$, $N_{NOM}N_{DAT}V_{psv}$, $N_{CASE}N_{CASE}V_{act}$, $N_{CASE}N_{CASE}V_{psv}$, $N_{CASE}V_{act}$, and $N_{CASE}V_{psv}$. This finding aligns with previous reports on language-specific challenges to the automatic processing of Korean (Shin, 2022b; Kim et al., 2007), also partially aligning with Ambridge et al. (2020) showing the failure of modelling human judgements in K'iche'.

Regarding language-specific and construction-specific properties, the models' capacity to recognise passive morphology and conduct the required revision process involving the suffixal passive did not emerge clearly. In $N_{CASE}N_{CASE}V_{psv}$ and $N_{CASE}V_{psv}$, the core conditions testing how the models cope with passive morphology and its related interpretive procedures for classification, not all the sub-models succeeded in classifying the test stimuli as *Theme-First* as intended ($N_{CASE}N_{CASE}V_{psv}$: Learning rate = 0.0001, Batch = 16, MaxSeqLen = 256; $N_{CASE}V_{psv}$: Learning rate = 0.0001). Moreover, the classification accuracy of model outputs in

$N_{NOM}N_{DAT}V_{psv}$, $N_{DAT}N_{NOM}V_{psv}$, $N_{NOM}V_{psv}$, and $N_{DAT}V_{psv}$ did not seem to reasonably approximate the children's picture-selection patterns found in Shin (2022a), also manifesting notable by-hyperparameter asymmetries. The precise locus of these asymmetries seems nebulas, as is often the case when interpreting LLMs' performance against downstream language tasks. However, the divergence between the models' performance and the children's comprehension behaviour in the suffixal passive conditions imply that neural networks are not agile with coping with linguistic cues that are language specific, or at least, neural networks handle linguistic cues differently from the (developing) human processor does so.

Another factor possibly contributing to the models' performance is the simulation environments in this study. We trained each model with all the transitive-event instances in CHILDES, considering how the children in Shin (2022a) attuned their interpretation to transitive events before being exposed to the stimuli. Despite this treatment, the models' testing environment may not have fully conformed to what the children partially experienced due to the pre-trained models,

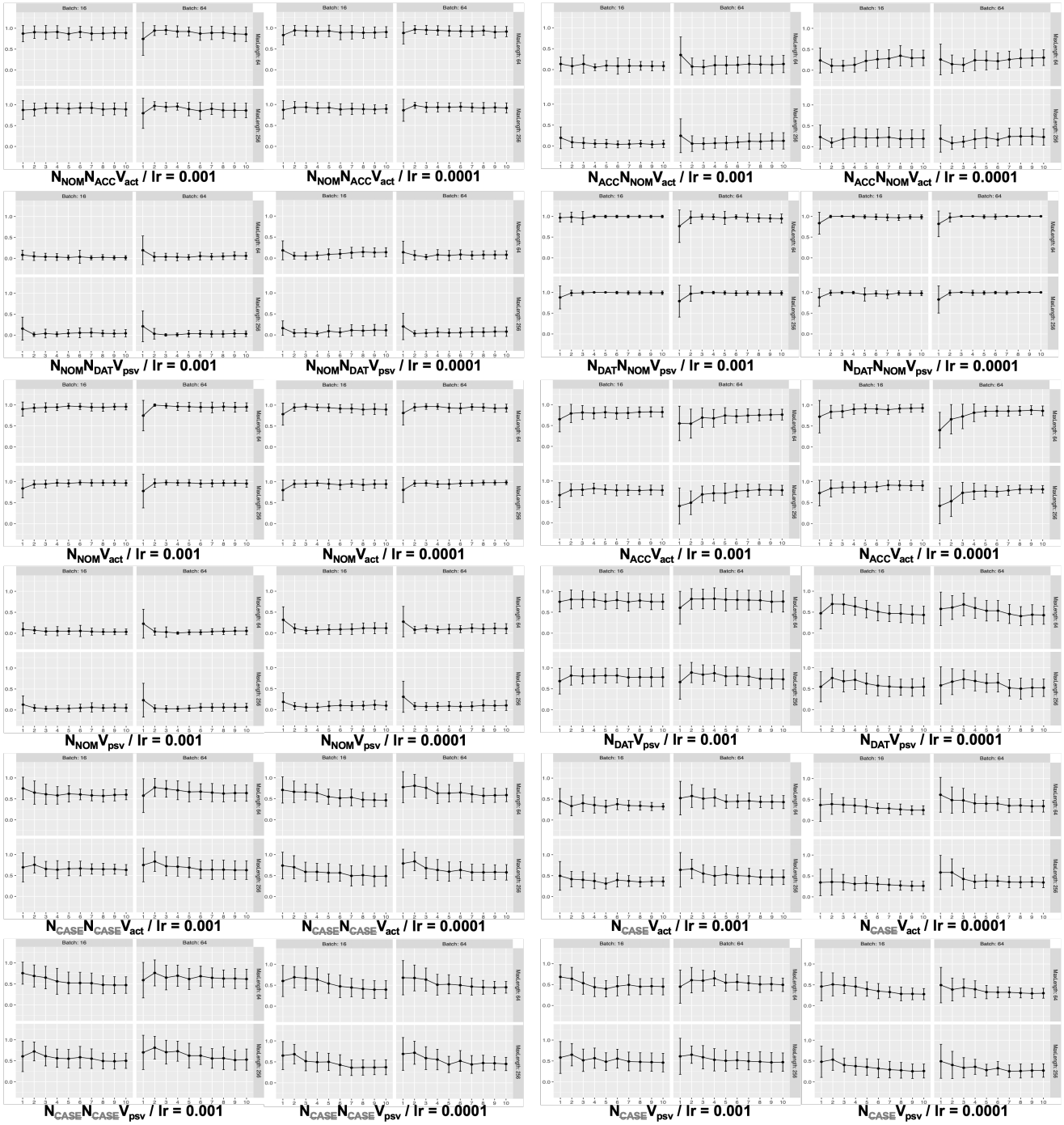

Figure 2. Model performance by condition and learning rate. X-axis = epoch; Y-axis = mean accuracy for case-marked conditions; mean rate of agent-first response for case-less conditions. Error bars = 95% CIs.

mostly comprising adult language features, when constructing each model. Moreover, the test items in the simulations involved no overt acoustic-masking effects as used in Shin (2022a) that informed the children of something that was somehow hidden. This absence of auditory signals about the marker(s), which was inevitable given the simulation settings in which the models worked exclusively with the textual data, may have affected the model performance unexpectedly (cf. Stoyneshka et al., 2010). Together, the simulations stood on a somewhat different ground than the experiments (as most modelling research does), possibly inviting the observed model–children asymmetry to the extent that the models did not handle the stimuli in the same way as the children did in the experiments. However, we cannot jump to a firm conclusion that these are the all-and-only reasons explaining the disparity between the models' performance and the children's picture-selection patterns in Shin (2022a) because these issues have not been fully explored in this field.

In addition to these factors, algorithmic characteristics of a computational architecture may be a core source of this disparity. Neural networks often utilise contextual information through window-based computation (Haykin, 2009; Kriesel, 2007) when presented with a sampling of data points. This involves driving contextual information from formal sequences of words or characters, relying heavily on form (cf. Firth, 1957) rather than a context in a linguistic sense. In other words, when the models access the meaning or function of a linguistic unit, they resort to the formal co-occurrences in the incoming input, not directly drawing upon the meaning or function of that unit. Moreover, while neural networks excel at generalising from pre-trained models and fine-tuning data, they struggle with extrapolating beyond their training space (Marcus, 1998). Deep-learning models attempt to resolve this issue by using massive amounts of data to cover every potential instance of formal co-occurrences; state-of-the-art LLMs with billions of parameters benefit from deploying exceedingly large training sets. They often yield good performance when handling known inputs but remain unsatisfactory with novel inputs (cf. Choi, 2023), particularly for accessing meaning or function through form (Ettinger et al., 2023; West et al., 2023). More broadly, computational models encounter language usage indirectly and not in a grounded manner; they do not directly engage in language-usage profiles and situations to which language refers (Clark, 1996; McClelland et al., 2020). Relatedly, neural networks are particularly susceptible to frequency effects (Marcus, 1995; McCurdy et al., 2020).

This nature may have caused the models' performance to deviate from the children's response patterns on some test items which could be out of range. The stimuli in Shin (2022a), consisting of animal names as entities, would be new instances for our models in this respect (and also considering the typical composition of transitive sentences in ordinary speech—animate agents and inanimate themes e.g., Dowty, 1991). Some stimuli involved scrambling or omission of sentential components, which are also non-typical. These factors may have led the models to malfunction in their operation. The key evidence for this argument comes from the models' performance on the conditions in which a simulated learner must determine the thematic role of the first and sole case-less noun only with its presence ($N_{\not{C}ASE}V_{act}$; $N_{\not{C}ASE}V_{psv}$) compared to their performance on one-argument case-marked conditions in which a simulated learner has more information about the first noun's thematic role indicated by a case marker next to the noun ($N_{NOM}V_{act}$; $N_{ACC}V_{act}$; $N_{NOM}V_{psv}$; $N_{DAT}V_{psv}$).

Relatedly, the notable variations in the models' performance generated by hyperparameter manipulation further support our claim regarding the major role of algorithmic characteristics of a computational model for simulating human language behaviour. We found that, of the three hyperparameters we selected in each architecture, the learning rate was the most influential in adjusting the models' classification behaviour. Considering its concept in machine learning (i.e., a hyperparameter that controls the rate at which an algorithm updates or learns the values of a parameter estimate), it likely serves as a proxy for the degree and manner to which humans generalise (linguistic) knowledge. Scholars have debated how learners derive linguistic knowledge from concrete items and apply it towards abstract representations—gradual abstraction (conservatism when transferring current knowledge to new items; Ambridge & Lieven, 2015; Theakston et al., 2015) vs. early abstraction (rapid generalisation of current knowledge to other relevant items; Fisher, 1996; Lidz et al., 2003). If our approach is on the right track, the simulations in this study could open a new window to complementing and advancing the literature on how children generalise linguistic knowledge as a function of exposure to linguistic environments and domain-general learning capacities. Nevertheless, we concede that our claim here is based on exploratory observations and is, therefore, speculative. Thus, further examination is needed.

Conclusion

Our study revealed that, while the GPT-2 architecture tested in this study may be able to utilise information about formal co-occurrences to access the intended message to a certain degree, (the outcome of) this process may substantially differ from how a child, as a developing processor, engages in comprehension of this construction. To discern verbal morphology indicating the voice and recalibrate the initial, garden-pathed alignments between thematic roles and case markers to formulate a correct interpretation, the child processor likely draws upon multiple morpho-syntactic and semantic cues, which are searchable from their language-usage profiles and are sensitive to usage frequencies. Moreover, its operation is likely influenced by multiple sources, including event/world knowledge (Friedman, 2000; Snedeker & Trueswell, 2004), memory operation (Kim et al., 2017), task type (Huang et al., 2013), and cognitive bias (e.g., *Agent-First* strategy; Shin, 2021; Abbot-Smith et al., 2017). This interplay may not have been properly captured and modelled by the modelled learners in this study. We believe this study provides evidence of the limits of the neural networks' capacity to address child language features. This invites subsequent inquiries on the extent to which computational models reveal developmental trajectories of child language that have been unveiled through corpus-based or experimental research.

References

Abbot-Smith, K., Chang, F., Rowland, C., Ferguson, H., & Pine, J. (2017). *Do two and three year old children use an incremental first-NP-as-agent bias to process active transitive and passive sentences?: A permutation analysis*. PLoS One, 12(10), e0186129. https://doi.org/10.1371/journal.pone.0186129

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2), 119–126. https://doi.org/10.1080/00031305.1998.10480550

Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science*, 32(5), 789–834. https://doi.org/10.1080/03640210801929287

Ambridge, B., & Lieven, E. (2015). A constructivist account of child language acquisition. In B. MacWhinney, & W. O'Grady (Eds.), *The Handbook of Language Emergence* (pp. 478–510). Malden, MA: John Wiley & Sons

Ambridge, B., Maitreyee, R., Tatsumi, T., Doherty, L., Zicherman, S., Pedro, P. M., Bannard, C., Samanta, S., McCauley, S., Arnon, I., Bekman, D., Efrati, A., Berman, R., Narasimhan, B., Sharma, D. M., Nair, R. B., Fukumura, K., Campbell, S., Pye, C., Pixabaj, S. F. C., Paliz, M. M., & Mendoza, M. J. (2020). The crosslinguistic acquisition of sentence structure: Computational modeling and grammaticality judgments from adult and child speakers of English, Japanese, Hindi, Hebrew and K'iche'. *Cognition*, 202, 104310. https://doi.org/10.1016/j.cognition.2020.104310

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, 106(41), 17284–17289. https://doi.org/10.1073/pnas.0905638106

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Budzianowski, P., & Vulić, I. (2019). Hello, it's GPT-2 - How can I help you? Towards the use of pretrained language models for task-oriented dialogue systems. In A. Birch, A. Finch, H. Hayashi, I. Konstas, T. Luong, G. Neubig, Y. Oda, & K. Sudoh (Eds.), *Proceedings of the 3rd Workshop on Neural Generation and Translation* (pp. 15–22). Association for Computational Linguistics. https://aclanthology.org/D19-5602

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, 27(6), 843–873. https://doi.org/10.1207/s15516709cog2706_2

Chang, F. (2002). Symbolically speaking: A connectionistmodel of sentence production. *Cognitive Science*, 26(5), 609–651. https://doi.org/10.1207/s15516709cog2605_3

Chang, F. (2009). Learning to order words: A connectionist model of heavy NP shift and accessibility effects in Japanese and English. *Journal of Memory and Language*, 61(3), 374–397. https://doi.org/10.1016/j.jml.2009.07.006

Choi, Y. (2023). Common Sense: The Dark Matter of Language and Intelligence (VLDB 2023 Keynote). *Proceedings of the VLDB Endowment*, 16(12), 4139–4139. https://doi.org/10.14778/3611540.3611638

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, 47(3), e13256. https://doi.org/10.1111/cogs.13256

Dai, D., Sun, Y., Dong, L., Hao, Y., Ma, S., Sui, Z., & Wei, F. (2023). Why can GPT learn in-context? Language models secretly perform gradient descent as meta-optimizers. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 4005–4019). Association for Computational Linguistics. https://aclanthology.org/2023.findings-acl.247

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67(3), 547–619. https://doi.org/10.1353/lan.1991.0021

Edwards, C. (2015). Growing pains for deep learning. *Communications of the ACM*, 58(7), 14–16. https://doi.org/10.1145/2771283

Ettinger, A., Hwang, J. D., Pyatkin, V., Bhagavatula, C., & Choi, Y. (2023). " You Are An Expert Linguistic Annotator": Limits of LLMs as analyzers of abstract meaning representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 8250–8263). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2310.17793

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–55. In *Studies in Linguistic Analysis* (pp. 1–31). Special Volume of the Philological Society. Oxford: Blackwell [Reprinted as Firth (1968)].

Friedman, W. J. (2000). The development of children's knowledge of the times of future events. *Child Development*, 71(4), 913–932. https://doi.org/10.1111/1467-8624.00199

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Reichart, R., Devore, S., Flinker, A., Hasenfratz, L., Levy, O., Hassidim, A., Brenner, M., Matias, Y., Norman, K. A., Devinsky, O., & Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25, 369–380. https://doi.org/10.1038/s41593-022-01026-4

Goldstein, A., Ham, E., Nastase, S. A., Zada, Z., Grinstein-Dabush, A., Aubrey, B., Schain, M., Gazula, H., Feder, A., Doyle, W., Devore, S., Dugan, P., Friedman, D., Brenner, M., Hassidim, A., Devinsky, O., Flinker, A., Levy, O., & Hasson, U. (2022). Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain. *BioRxiv*. https://doi.org/10.1101/2022.07.11.499562

Hawkins, R. D., Yamakoshi, T., Griffiths, T. L., & Goldberg, A. E. (2020). Investigating representations of verb bias in neural language models. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 4653–4663). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2010.02375

Haykin, S. (2009). *Neural networks and learning machines*. Prentice Hall.

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2022). Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. *BioRxiv*, 2022-10. https://doi.org/10.1101/2022.10.04.510681

Huang, Y. T., Zheng, X., Meng, X., & Snedeker, J. (2013). Children's assignment of grammatical roles in the online processing of Mandarin passive sentences. *Journal of Memory and Language*, 69(4), 589–606. https://doi.org/10.1016/j.jml.2013.08.002

Fisher, C. (1996). Structural limits on verb mapping: the role of analogy in children's interpretation of sentences. *Cognitive Psychology*, 31, 41–81. https://doi.org/10.1006/cogp.1996.0012

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). A systematic assessment of syntactic generalization in neural language models. *The Proceedings of the Association for Computational Linguistics*. https://doi.org/10.48550/arXiv.2005.03692

Kendeou, P., Smith, E. R., & O'Brien, E. J. (2013). Updating during reading comprehension: why causality matters. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(3), 854–865. https://doi.org/10.1037/a0029468

Kim, B., Lee, Y., & Lee, J. (2007). Unsupervised semantic role labeling for Korean adverbial case. *Journal of KIISE: Software and Applications*, 34(2), 32–39.

Kim, J-B., & Choi, I. (2004). The Korean case system: A unified, constraint-based approach. *Language Research*, 40, 885–921.

Kim, S. Y., Sung, J. E., & Yim, D. (2017). Sentence comprehension ability and working memory capacity as a function of syntactic structure and canonicity in 5-and 6-year-old children. *Communication Sciences & Disorders*, 22(4), 643–656. https://doi.org/10.12963/csd.17420

Kriesel, D. (2007). A brief introduction to neural networks. Available at http://www.dkriesel.com (accessed on 2023-11-07)

Lidz, J., Gleitman, H., & Gleitman, L. (2003). Understanding how input matters: Verb learning and the footprint of universal grammar. *Cognition*, 87(3), 151–178. https://doi.org/10.1016/S0010-0277(02)00230-5

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed). Lawrence Erlbaum.

Marcus, G. F. (1995). The acquisition of the English past tense in children and multilayered connectionist networks. *Cognition*, *56*(3), 271–279. https://doi.org/10.1016/0010-0277(94)00656-6

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, 37(3), 243–282. https://doi.org/10.1006/cogp.1998.0694

Marvin, R., & Linzen, T. (2019). Targeted syntactic evaluation of language models. *Proceedings of the Society for Computation in Linguistics*, 2(1), 373–374. https://doi.org/10.48550/arXiv.1808.09031

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, 117(42), 25966–25974. https://doi.org/10.1073/pnas.1910416117

McCurdy, K., Goldwater, S., & Lopez, A. (2020). Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1745–1756). https://doi.org/10.18653/v1/2020.acl-main.159

O'Grady, W., & Lee, M. (2023). Natural Syntax, Artificial Intelligence and language acquisition. *Information*, 14(7), 418. https://doi.org/10.3390/info14070418

Oh, B. D., Clark, C., & Schuler, W. (2022). Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5, 777963. https://doi.org/10.3389/frai.2022.777963

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, 118(3), 306–338. https://doi.org/10.1016/j.cognition.2010.11.001

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 9.

Rapp, D. N., & Kendeou, P. (2007). Revising what readers know: Updating text representations during narrative comprehension. *Memory & Cognition*, 35(8), 2019–2032. https://doi.org/10.3758/BF03192934

Sagae, K. (2021). Tracking child language development with neural network language models. *Frontiers in Psychology*, 12, 674402. https://doi.org/10.3389/fpsyg.2021.674402

Shin, G-H. (2021). Limits on the Agent-First strategy: Evidence from children's comprehension of a transitive construction in Korean. *Cognitive Science, 45*(9), e13038. https://doi.org/10.1111/cogs.13038

Shin, G-H. (2022a). Awareness is one thing and mastery is another: Korean-speaking children's comprehension of a suffixal passive construction in Korean. *Cognitive Development, 62,* 101184. https://doi.org/10.1016/j.cogdev.2022.101184

Shin, G-H. (2022b). Automatic analysis of caregiver input and child production: Insight into corpus-based research on child language development in Korean. *Korean Linguistics, 18*(2), 125–158. https://doi.org/10.1075/kl.20002.shi

Shin, G-H., & Deen, K. (2023). One is not enough: Interactive role of word order, case marking, and verbal morphology in children's comprehension of suffixal passive in Korean. *Language Learning and Development, 19*(2), 188–212. https://doi.org/10.1080/15475441.2022.2050237

Shin, G-H., & Mun, S. (2023a). Korean-speaking children's constructional knowledge about a transitive event: Corpus analysis and Bayesian modelling. *Journal of Child Language, 50*(2), 311–337. https://doi.org/10.1017/S030500092100088X

Shin, G-H. & Mun, S. (2023b). Explainability of neural networks for child language: Agent-First strategy in comprehension of Korean active transitive construction. *Developmental Science*, e13405. https://doi.org/10.1111/desc.13405

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, 49(3), 238–299. https://doi.org/10.1016/j.cogpsych.2004.03.001

Sohn, H. M. (1999). *The Korean language*. Cambridge University Press.

Stoll, S., Abbot-Smith, K., & Lieven, E. (2009). Lexically restricted utterances in Russian, German, and English child-directed speech. *Cognitive Science*, 33(1), 75–103. https://doi.org/10.1111/j.1551-6709.2008.01004.x

Stoyneshka, I., Fodor, J. D., & Fernández, E.M. (2010). Phoneme restoration methods for investigating prosodic influences on syntactic processing. *Language and Cognitive Processes*, 25(7-9), 1265–1293. https://doi.org/10.1080/01690961003661192

Theakston, A. L., Ibbotson, P., Freudenthal, D., Lieven, E. V., & Tomasello, M. (2015). Productivity of noun slots in verb frames. *Cognitive Science*, 39(6), 1369–1395. https://doi.org/10.1111/cogs.12216

Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten-path effect: studying on-line sentence processing in young children. *Cognition*, 73, 89–134. https://doi.org/10.1016/S0010-0277(99)00032-3

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Proceedings of the 31st advances in Neural Information Processing Systems* (pp. 5998–6008). Curran Associates, Inc.

de Vries, W., & Nissim, M. (2021). As good as new. How to successfully recycle English GPT-2 to make models for other languages. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 836–846). Association for Computational Linguistics. https://aclanthology.org/2021.findings-acl.74

Warstadt, A., & Bowman, S. R. (2020). Can neural networks acquire a structural bias from raw linguistic data? In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 1737–1743) Cognitive Science Society. https://doi.org/10.48550/arXiv.2007.06761

Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7, 625–641. https://doi.org/10.1162/tacl_a_00290

West, P., Lu, X., Dziri, N., Brahman, F., Li, L., Hwang, J. D., Jiang, L., Fisher, J., Ravichander, A., Chandu, K., & Newman, B. (2023). The Generative AI paradox:" What it can create, it may not understand". arXiv preprint. https://doi.org/10.48550/arXiv.2311.00059

Xu, W., Chon, J., Liu, T., & Futrell, R. (2023). The Linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 15711–15721). https://doi.org/10.18653/v1/2023.findings-emnlp.1052

You, G., Bickel, B., Daum, M. M., & Stoll, S. (2021). Child-directed speech is optimized for syntax-free semantic inference. *Scientific Reports*, 11(1), 1–11. https://doi.org/10.1038/s41598-021-95392-x