Korean Journal of English Language and Linguistics, Vol 22, September 2022, pp. 1033-1050
DOI: 10.15738/kjell.22..202210.1033

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

(AL)BERT Down the Garden Path: Psycholinguistic Experiments for Pre-trained Language Models*

Jonghyun Lee (Seoul National University) **Jeong-Ah Shin** (Dongguk University)
Myung-Kwan Park (Dongguk University)

Received: September 03, 2022
Revised: September 25, 2022
Accepted: September 30, 2022

Jonghyun Lee (first author)
Graduate Student (PhD), Dept. of English Language and Literature,
Seoul National University
Email: museeq@snu.ac.kr

Jeong-Ah Shin (corresponding author)
Professor, Division of English Language and Literature, Dongguk University
Email: jashin@dongguk.edu

Myung-Kwan Park (co-author)
Professor, Division of English Language and Literature, Dongguk University
Email: korgen2003@naver.com

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A03031616).

ABSTRACT

Lee, Jonghyun, Jeong-Ah Shin and Myung-Kwan Park. 2022. (AL)BERT Down the Garden Path: Psycholinguistic experiments for pre-trained language models. *Korean Journal of English Language and Linguistics* 22, 1033-1050.

This study compared the syntactic capabilities of several neural language models (LMs) including Transformers (BERT / ALBERT) and LSTM and investigated whether they exhibit human-like syntactic representations through a targeted evaluation approach, a method to evaluate the syntactic processing ability of LMs using sentences designed for psycholinguistic experiments. By employing garden-path structures with several linguistic manipulations, whether LMs detect temporary ungrammaticality and use a linguistic cue such as plausibility, transitivity, and morphology is assessed. The results showed that both Transformers and LSTM exploited several linguistic cues for incremental syntactic processing, comparable to human syntactic processing. They differed, however, in terms of whether and how they use each linguistic cue. Overall, Transformers had a more human-like syntactic representation than LSTM, given their higher sensitivity to plausibility and ability to retain information from previous words. Meanwhile, the number of parameters does not seem to undermine the performance of LMs, contrary to what was predicted in previous studies. Through these findings, this research sought to contribute to a greater understanding of the syntactic processing of neural language models as well as human language processing.

KEYWORDS

targeted evaluation approach, transformers, garden-path structure, natural language processing, psycholinguistics

1. Introduction

Pre-trained neural language models using the Transformer network such as Bidirectional Encoder Representations from Transformers (BERT, Devlin et al. 2019) and A Lite BERT (ALBERT, Lan et al. 2020) have recently shown superior performance on several language understanding benchmarks. They seem to learn robust linguistic competence about natural language during the pre-training process, which enables them to be applied to task-specific fine-tuning while maintaining strong performance (Ettinger 2020). However, it is not yet fully understood whether and how these models are successful in acquiring those capacities, including syntactic generalizations, during the pre-training process.

This paper examines human-like, generalizable syntactic competence of the Transformers such as BERT and ALBERT, comparing them with more traditional Long Short-Term Memory (LSTM) language model (Gulordava et al. 2018) in order to enhance the understanding of the pre-training NLP models. For this purpose, a targeted evaluation approach (Marvin and Linzen 2018), or psycholinguistic experimental techniques, was employed. This approach, introduced by Linzen, Dupoux and Goldberg (2016), is the way of diagnosing a language model's (LM) syntactic language capacity by observing its behavior when it processes carefully constructed sentences that are required to rely on particular syntactic representations. As in human language processing, examining LM's language processing will help to study LM's capacities on several aspects of language.

1.1 Previous Studies

Many studies have been conducted using a targeted evaluation approach to explore the neural language models' syntactic representations. They examined, mostly using RNN models as LM, various grammatical aspects of language processing such as subject-verb agreement, long distance dependency, and anaphoric dependency (e.g., Marvin and Linzen 2018). The results revealed that LMs had robust syntactic generalization competence, but did not always achieve human-like syntactic performance. Rather, the experimental results were mixed depending on sentence structures targeted or used in the models. For example, when Futrell et al. (2019) tested four different LSTM models on several structures, two of the models fully showed human-like syntactic behavior, whereas the others did not.

Although much of previous research has mainly focused on RNN (or LSTM), some recent works investigated the Transformer architectures such as BERT and ELMO to see if these pre-trained models also captured robust syntactic generalization. In Goldberg (2019), BERT exhibited strong performance on subject-verb agreement syntactic tasks. Ettinger (2020) also showed that BERT successfully retrieved a correct word in the word completion task, although it was less sensitive than humans and had some struggles with challenging inference and role-based event prediction. Several studies have also compared the Transformer models with RNN in terms of human-like syntactic representations (Hu et al. 2020, van Schijndel, Mueller and Linzen 2019, Wilcox, Levy and Futrell 2019). Some of them suggested that Transformer models did not show more human-like performance in several syntactic tests such as subordination parsing (Hu et al. 2020), subject-verb agreement (van Schijndel, Mueller and Linzen 2019) and filler–gap dependency (Wilcox, Levy and Futrell 2019), even though they generally outperformed RNN in the language understanding benchmarks. From these results, Wilcox, Levy and Futrell (2019) suggested that the larger number of parameters in the Transformer may not necessarily contribute to more human-like syntactic performance. On the other hand, in Hu et al. (2020), GPT-2-XL, the largest Transformer among tested models, scored the highest accuracy in the syntactic tests.

1.2 Present Study

The current study tested the syntactic capacities of pre-trained LMs such as ALBERT, BERT and LSTM through a targeted evaluation approach as in previous studies. However, this study differs from previous studies largely in two aspects. First, the syntactic tests were conducted not only with BERT and LSTM, but also with the relatively recent model, ALBERT. ALBERT, one of Transformer architecture, is similar to BERT, but two parameter reduction techniques are applied to reduce longer training time and excessive memory consumption, which were the shortcomings of BERT (Lan et al. 2020). The first technique is a factorized embedding parameterization, which is to decompose the large word embedding matrix into two smaller matrices. The second technique is cross-layer parameter sharing, which increases efficiency by sharing parameters across layers. This is also used by other Transformers, but they share either feed-forward network parameters or attention parameters, whereas ALBERT shares all parameters. Through these techniques, Lan et al. (2020) has improved the model to have similar or superior language competence while reducing training time and memory consumption. This study included ALBERT, an improved Transformer model, for testing. Second, by varying the number of parameters in the model, it is attempted to determine whether the less human-like syntactic performance of the Transformer models is attributed to a large number of parameters, as mentioned in previous studies (van Schijndel, Mueller and Linzen 2019, Wilcox, Levy and Futrell 2019). As for ALBERT, four sub-models with different number of parameters are provided. For example, albert-base, the smallest sub-model, has 12M parameters, and albert-xxlarge, the largest, has 235M parameters (See Table 1 for the information of the other models). The present study examined how the LM's syntactic generalization competence varies according to the number of parameters.

This study focuses on neural language models' performance on incremental syntactic processing, using temporarily ungrammatical sentences, such as garden-path structures. Much psycholinguistics research has examined syntactic states in incremental processing as a method to measure human syntactic knowledge (Futrell et al. 2018). This is because it can provide a wealth of information on how comprehenders, in real time, syntactically respond to words they encounter and make grammatical predictions for words to come. As the current study aims to explore whether LMs have syntactic representations comparable to humans, as in the study of psycholinguistics, it investigates how LMs represent the currently unfolding word and predicts the upcoming word during incremental syntactic processing. This study employed temporally ungrammatical sentences, or garden-path structure, to examine the incremental syntactic states of LMs. Garden-path structures are grammatical in the whole context of the sentence, but they can be judged as temporarily ungrammatical at the position of a particular word when processed in real time. These sentences have an advantage in that a test can be constructed with only grammatically licensed sentences. The sentences with clear, right or wrong, grammatical judgment, such as subject-verb agreement structure, must include grammatically unlicensed sentences in the test suite. Given that the majority of training sentences provided for LMs are grammatically licensed sentences, this test set is more suitable to investigate the syntactic representation that LMs processed based on the input they received during the training. In addition, garden-path structures are conducive to explore complicated semantic and syntactic interactions between subject, verb, and object. A change in the plausibility between a verb and an object, for instance, might alter the comprehension of the grammatical relation between a verb and an object (See 2.3 for more concrete examples). By adjusting language cues such as plausibility, garden-path structures allow the investigation of diverse relationships between grammatical elements.

In this regard, four research questions were addressed: (1) Do deep neural network language models such as Transformers and LSTM show incremental syntactic representation comparable with human language processing? (2) What linguistic cues influence the LMs' syntactic processing? (3) How is the syntactic representation of more recent models such as ALBERT and BERT different from more traditional LSTM models in terms of human-like syntactic processing? (4) Does the different number of parameters of pre-trained models affect their syntactic processing?

2. Method

2.1 Dependent Measures

An LM's syntactic processing is measured by the surprisal,

$$S(x_i) = -\log_2 p(x_i|h_{i-1}),$$

or the log inverse probability of a target word, x_i, given the previous hidden state, h_{i-1}. The probability is calculated from the LM's softmax activation (Futrell et al., 2018). In the LSTM model, the previous hidden state is the segment of the sentence before consuming the target word (e.g. "As the woman edited the magazine (amused all the reporters.)" for the target word, amused), while in the BERT and ALBERT models, it is the masked sentence with a [MASK] token, in the position of the target word (e.g. "[CLS] As the woman edited the magazine [MASK] all the reporters. [SEP]" for the target word, pleased). Since BERT and ALBERT are bidirectional models which are able to and required to use information from both left and right direction, the contexts after the target word are also presented, unlike the LSTM model which uses unidirectional information.

In psycholinguistics, the reading times (RTs) for a word are considered to reflect humans' expectation on a particular word during incremental sentence processing (Jegerski 2013, Just and Carpenter 1980). In general, when ones encounter unexpected words in a given context, their RTs increase. Since ungrammaticality—temporary or not—is unexpected, it generally increases RTs, and this increase is used as evidence of human behavior of syntactic state representations. These RTs are known to be generally proportional to the surprisal of the probabilistic language model of the comprehender (Levy 2008, Smith and Levy 2013). It has been also known that LMs' surprisal is a strong predictor of human reading times (Demberg and Keller 2008, Goodkind and Bicknell 2018). In this study, following Futrell et al. (2018), LMs' surprisal is considered as analogue to human reading times, and used as a dependent measure to examine LM's prediction of the target word.

2.2 Neural Language Models

The neural language models tested in the study are BERT (Devlin et al. 2019), ALBERT (Lan et al. 2020) and LSTM (Gulordava et al. 2018). BERT is a deep bidirectional transformer network, pre-trained from masked texts by jointly conditioning on both left and right context in all layers (Devlin et al. 2019). Two versions of BERT are tested—bert-base-uncased and bert-large-uncased. They have the same basic architecture with the different parameters. ALBERT (A lite BERT) is a light version of BERT, which reduces longer training time and larger memory consumption of BERT. Four versions of ALBERT were tested—albert-base-v2, albert-large-v2, albert-xlarge-v2, and albert-xxlarge-v2. They are different in the number of parameters. Hugging Face[1] implementation was used as the pre-trained models for BERT and ALBERT in the experiments[2]. Hugging Face provides pre-trained models of various Transformers such as BERT. By loading the pre-trained models from it, it is possible to

[1] https://huggingface.co/transformers/pretrained_models.html

[2] Codes for the experiments will be provided upon request.

test the models without time-consuming training procedure.

The models are summarized in Table 1. As for BERT and ALBERT, the experimental sentences are processed to have a [MASK] token in the position of the target words, whose surprisal is measured for the test. Following Goldberg (2019) and Ettinger (2020), a [CLS] token is inserted at the beginning of each sentence to simulate the training conditions of the model. A [SEP] token is also included after the end of each sentence to indicate the end of the sentence to the models. The LSTM model tested (ColorlessgreenRNN) is adapted from Gulordava et al. (2018), which has been frequently tested in the previous studies using a targeted evaluation approach. The model is trained on 90 million tokens of English Wikipedia with two hidden layer dimensionalities (650 and 200 units).

Table 1. The Configurations of the BERT and ALBERT Models Tested

Model		Vocab Size	Parameters	Layers	Hidden	Embedding
BERT	base	30522	108M	12	768	768
	large	30522	334M	24	1024	1024
ALBERT	base	30000	12M	12	768	128
	large	30000	18M	24	1024	128
	xlarge	30000	60 M	24	2048	128
	xxlarge	30000	235M	12	4096	128
LSTM		50001	39M	2	200	650

2.3 Target Sentence Structures

Three types of garden-path structures were tested, which differ in linguistic cues: (1) plausibility, (2) transitivity and (3) morphology (See Results for the example sentence of each test set). Each sentence structure consists of garden-path structures that induce temporary ungrammaticality and have different linguistic cues. Garden-path sentences are grammatically correct sentences, but likely to be temporarily misinterpreted as ungrammatical during online processing (Bever, 1970). “As the woman edited the magazine amused all the reporters” is an example of a typical garden-path structure, so-called subject-object ambiguities sentence (Trueswell, Tanenhaus and Garnsey 1994). Although this sentence is grammatical, during incremental processing, *the magazine*, which is in fact the subject of the main verb, *amused*, may be misinterpreted as the object of *edited*. This misinterpretation causes comprehenders to temporarily believe the sentence is ungrammatical, that is, there is no subject of the main verb, when they encounter *amused*. Here, *amused* is referred to as a disambiguating word, since this is the position where the ambiguity of interpretation—whether to regard *the magazine* as the object of *edited* or as the subject of *amused*—is identified and possibly resolved. It is known that reading times (or surprisal) at the disambiguating position increase if comprehenders detect the temporary ungrammaticality and reanalyze the sentence structure (Frazier and Rayner 1982), which is known as garden-path effect.

However, the garden-path effect does not always occur, even with the aforementioned typical structure. Depending on the relationship between the two words causing the effect, it may diminish or disappear. In the above example, if the verb and the following noun are not closely related, the following word will not be interpreted as the object of the preceding verb; hence, there is no temporary ungrammaticality at the disambiguating position. This study referred these factors that influence the interaction between two words as linguistic cues and attempted to vary them to evaluate LMs’ syntactic performance. These cues include plausibility, transitivity, and morphology, and each will reveal which language aspects the LMs employ effectively in syntactic processing and which they do not. To summarize, the target sentences, garden-path structures, will show the following results: (1) whether an LM shows the garden-path effect, or whether it detects the temporary ungrammaticality at the disambiguating word and (2) what linguistic cues an LM processes with, and whether it is as sensitive as humans’ syntactic processing.

2.4 Statistical Analysis

To statistically confirm the differences of surprisal across the conditions, statistical analysis was conducted with linear mixed effect model (“lme4”; Baayen, Davidson and Bates 2008), using lmerTest package for R (Kuznetsova, Brockhoff and Christensen 2017). The lmerTest package calculates p-values for F-statistics *anova* and t-statistics *summary* of lme4 package (lmer model fits), with the Satterthwaite's degrees of freedom method. The p-values for fixed effect were obtained as follows: the mixed models were fitted using restricted maximum likelihood (REML) and the estimates of fixed effects and t-statistics were retrieved with the lme4 summary function. Then, the p-values for t-statistics were calculated using the lmerTest. In the results section, when the estimated p-value reaches a significant level, 0.05, it will be reported as a main effect or significant interaction of variables.

The mixed models for the analysis assume Garden-path (Garden-path *vs.* No garden-path), Linguistic cues (Plausibility, Transitivity, Morphology), and Length (Short *vs.* Long) as fixed effects, which makes the model a 2×2×2 analysis. Items (words) were included as random effects (random intercept) in order to minimize the influence of by-item variation. When an interaction between Length and the other effects were found, the separate analysis each for short and long sentence version was conducted to investigate the influence of intervening words on the garden-path effect, if necessary. The separate models include Garden-path and Linguistic cues as fixed effects. With this design, it will be considered that an LM detects the temporary ungrammaticality at the disambiguating word position if the main effect of Garde path occurs. Meanwhile, a significant 2×2 interaction (Garden-path × Linguistic cues) will suggest that an LM is sensitive to or influenced by that linguistic cue. A main effect of Length and an interaction between Length and the other variables will indicate the intervening words have influence on LMs processing of the garden-path sentences. In order to compare the differences between model architectures, Architecture (ALBERT, BERT, and LSTM) were included as factors in the statistical analysis, if necessary. Statistical analysis between architectures based on the difference in surprisal is difficult to produce reliable comparisons because the vocabulary used in each architecture and the number of them are different. Thus, statistical analysis was performed only when the opposite trend was observed between the architectures. Finally, a statistical analysis including Parameters as factors was also performed to explore the effect of the number of parameters on the syntactic states of the models. In this analysis, the number of each parameter was treated as continuous variables (e.g., 12 for albert-base); only four sub-models of ALBERT were analyzed, because each architecture was heterogeneous in terms of vocabulary size and average surprisal.

3. Results

3.1 Plausibility

For the first test, 24 items were adopted from Trueswell, Tanenhaus and Garnsey (1994) (see Example 1 below). They are the garden-path structure sentences that are assumed to induce subject-object ambiguities. Comprehenders might misinterpret *the magazine*, the subject of *amused*, as the object of *edited / sailed* while reading sentences such as (1a) and (1b). This will increase their RTs/surprisals at the point of the disambiguating word, *amused*. However, (1a) and (1b) differ in Plausibility. The verb phrase, *edited the magazine*, is plausible so that *the magazine* is more likely to be misunderstood as the object of the preceding verb, compared to *sailed the magazine*, the implausible one. (1c) and (1d) are No garden-path control sentences where *the magazine* is not ambiguous due to the presence of comma. The long version of the test set was also created by adding some intervening words—*about fishing* in Example (1)—between the noun and the disambiguating verb. This is to see if processing several intervening words before encountering disambiguating word will affect LMs' syntactic representation. In the psycholinguistics experiment with the similar materials, human comprehenders showed longer total reading times and more regressions at the disambiguating word in the plausible condition (1a) than the implausible conditions (1b) (Pickering & Traxler, 1998). In contrast, Roberts and Felser (2011) using a sentence slightly different from the one below did not find a difference in reading times by plausibility from English native speakers.

(1) a. As the woman edited the magazine (about fishing) amused all the reporters. [Plausible, Garden-path]
b. As the woman sailed the magazine (about fishing) amused all the reporters. [Implausible, Garden-path]
c. As the woman edited, the magazine (about fishing) amused all the reporters. [Plausible, No Garden-path]
d. As the woman sailed, the magazine (about fishing) amused all the reporters. [Implausible, No Garden-path]

Figure 1 shows the mean Garden-path effects (mean surprisal of Garden-path conditions *minus* No garden-path conditions) at the disambiguating word for all seven models, for both plausible and implausible conditions. Visual inspection on Figure 1 reveals garden-path effects for all LMs. However, overall pattern differences between Transformers and LSTM are also noticeable, despite no consistency across individual models. For BERT and ALBERT, the garden-path effect in the implausible conditions appears to be smaller than in the plausible conditions, while there seems to be no difference between two conditions in LSTM. In addition, a difference is observed in the effect by intervening words. The garden-path effect in BERT and ALBERT seems to be larger in long conditions, whereas smaller in LSTM. The statistical analysis for this visual inspection is presented below.

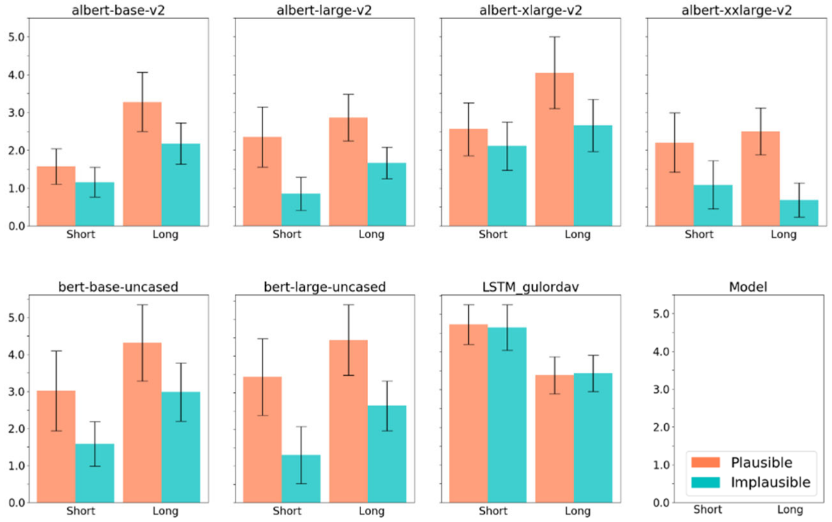


Figure 1. Mean Garden-path Effects (Mean Surprisal of the Garden-path Conditions *minus* No Garden-path Conditions) by Plausibility and Length across LMs

First, there was a main effect of Garden-path for all LMs including ALBERT, BERT and LSMT (albert-base: *estimate* = 2.05, *SE* = 0.29, *t* = 6.96, *p* < 0.001; albert-large: *estimate* = 1.93, *SE* = 0.36, *t* = 5.36, *p* < 0.001; albert-xlarge: *estimate* = 2.84, *SE* = 0.37, *t* = 7.62, *p* < 0.001; albert-xxlarge: *estimate* = 1.62, *SE* = 0.35, *t* = 4.59, *p* < 0.001; bert-base: *estimate* = 2.97, *SE* = 0.43, *t* = 6.89, *p* < 0.001; bert-large: *estimate* = 2.95, *SE* = 0.45, *t* =

6.57, $p < 0.001$; LSTM: *estimate* = 4.05, *SE* = 0.27, t = 14.74, $p < 0.001$). All LMs showed significantly larger surprisal at the disambiguating word in garden-path than no garden-path conditions. This suggests that all the LMs utilize the presence of comma as a cue of the clause boundary, or temporarily parse *the magazine* as the object of the preceding verb and then detect temporary ungrammaticality at the disambiguating word.

However, they did not reveal the interaction between Garden-path and Plausibility, except two models—albert-xxlarge (*estimate* = 1.47, *SE* = 0.71, t = 2.08, $p < 0.05$) and bert-large (*estimate* = 1.97, *SE* = 0.90, t = 2.19, $p < 0.05$). These two models exhibited a significant interaction between Garden-path and Plausibility, which is accounted for by larger garden-path effect for the plausible conditions compared to the implausible conditions. It might indicate that these models were less likely to misinterpret the following implausible noun of the verb as an object, using plausibility as a linguistic cue. Besides, a main effect of Plausibility was found for albert-large (*estimate* = 0.79, *SE* = 0.36, t = 2.20, $p < 0.05$) and bert-base (*estimate* = 0.96, *SE* = 0.43, t = 2.24, $p < 0.05$). Although this appears to be driven by greater surprisal of the plausible conditions within the garden-path condition, no significant interaction of Plausibility and Garden-path was found in the two models (marginal interaction for ablert-large ($p < 0.1$).

As for the comparison between short and long sentences, a main effect of Length was found only for albert-large (*estimate* = -0.81, *SE* = 0.43, t = 2.24, $p < 0.05$) and LSTM (*estimate* = -1.57, *SE* = 0.43, t = 2.24, $p < 0.001$). Their surprisal values were significantly larger in longer sentences than in shorter ones. In addition, there was a significant interaction between Length and Garden-path for albert-base (*estimate* = -1.36, *SE* = 0.59, t = -2.32, $p < 0.05$) and LSTM (*estimate* = 1.29, *SE* = 0.55, t = 2.34, $p < 0.001$) (Figure 2). In albert-base, garden-path effect was larger in the long condition, but in the LSTM, conversely, larger in the short condition, which is a reverse trend to what is generally expected of human language processing (Ferreira and Henderson 1991).

As for differences across the architectures, differences in garden-path effect by Plausibility as well as by Length were observed from descriptive statistics (Figure 3). However, only interaction among Length, Gardenpath and Architecture was significant (ALBERT *vs* LSTM: *estimate* = 2.03, *SE* = 1.02, t = 2.00, $p < 0.05$; BERT *vs* LSTM: *estimate* = 2.55, *SE* = 1.12, t = 2.28, $p < 0.05$). In ALBERT and BERT, the garden-path effect was greater when words were added, whereas in LSTM it was smaller. No effect was found with respect to Parameters.

To summarize, a simple garden-path effect was found for all LMs, but the difference in Plausibility was not statistically significant for all models. Nevertheless, the findings demonstrated that some of Transformers, bert-large and albert-xxlarge, which were LMs with the highest number of parameters, utilized plausibility cues when processing the subject-object ambiguity structure. Overall, it seems that Transformer models are overall more sensitive to plausibility than LSTM, but it was not statistically confirmed. The distinction between LSTM and Transform was derived from sentences in which syntactic processing became more complicated due to the addition of words. In contrast to Transformers, the garden-path effect of LSTM is diminished in long sentence conditions.

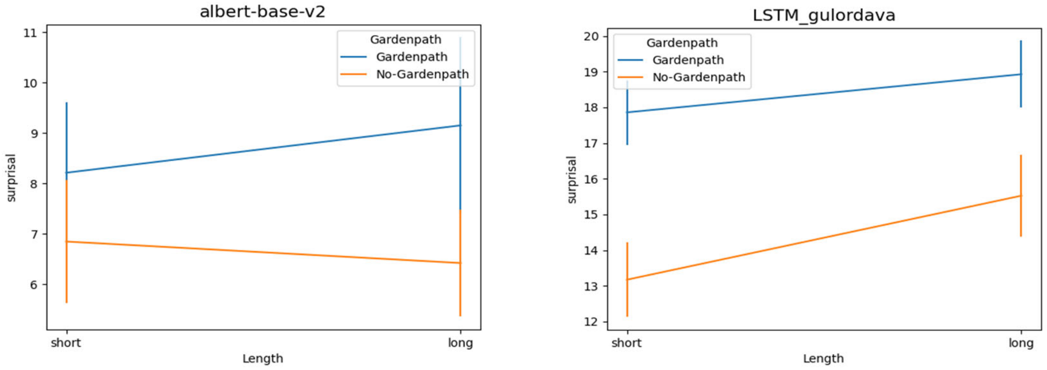
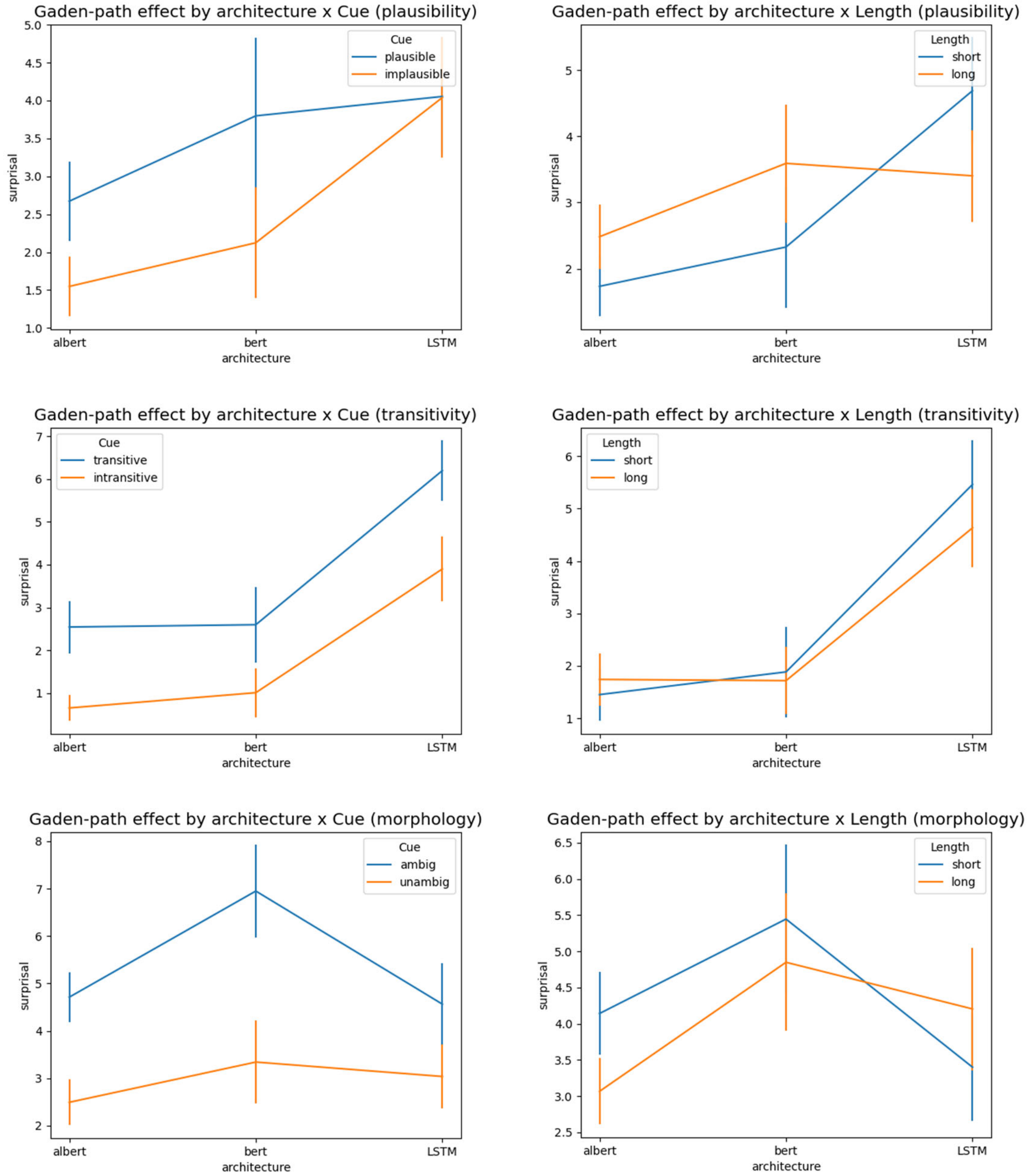


Figure 2. Mean Surprisal by Garden-path and Length for Albert-base (Left) and LSTM (Right)



Gaden-path effect by architecture x Cue (plausibility)

Gaden-path effect by architecture x Length (plausibility)

Gaden-path effect by architecture x Cue (transitivity)

Gaden-path effect by architecture x Length (transitivity)

Gaden-path effect by architecture x Cue (morphology)

Gaden-path effect by architecture x Length (morphology)

Figure 3. Mean Garden-path Effects by Cue x Architecture (Left) and Length x Architecture (Right) for Plausibility (Top), Transitivity (Middle) and Morphology (Bottom)

3.2 Transitivity

The second set of the sentences (24 items) was also a subject-object ambiguities structure such as Example (2), adopted from Futrell et al. (2019) and Staub (2007). Similarly with Example (1), comprehenders might initially assume *the vet* is the object of *scratched / struggled* in (2a) and (2b), increasing the RT/surprisal when they encounter the main verb phrase, *took off*, compared to the No garden-path sentences such as (2c) and (2d) where comma marks the end of the clause. However, while *scratched* in (2a) is a transitive verb that accepts an object, *struggle* in (2b) is an intransitive verb so that it is in principle not possible to interpret *the vet* as its object. Therefore, it is predicted that if LMs are sensitive to transitivity, larger surprisal will be found in (2a), compared to (2b). Several psycholinguistics experiments showed that humans take longer reading time in the transitive conditions than in the intransitive conditions (Adams, Clifton and Mitchell 1998, Staub 2007, van Gompel and Pickering 2001). Some intervening words such as *with his new assistant* were added in this test as well.

(2) a. When the dog scratched the vet (with his new assistant) took off the muzzle. [Transitive, Garden-path
b. When the dog struggled the vet (with his new assistant) took off the muzzle. [Intransitive, Garden-path]
c. When the dog scratched, the vet (with his new assistant) took off the muzzle. [Transitive, No garden-path]
d. When the dog struggled, the vet (with his new assistant) took off the muzzle. [Intransitive, No garden-path]

Figure 4 illustrates the average garden-path effects at the disambiguating word in both transitive and intransitive conditions across LMs. All models show a similar pattern, larger garden-path effects for transitive conditions.

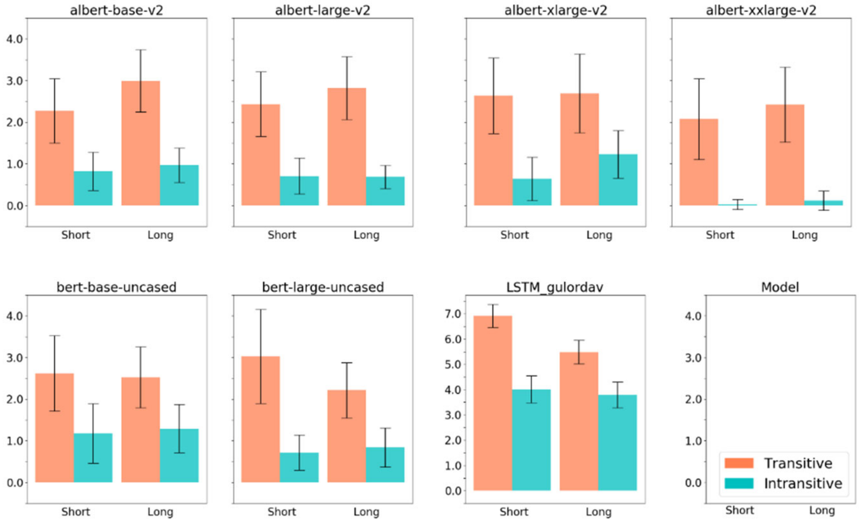


Figure 4. Mean Garden-path Effects by Transitivity and Length across LMs

The results revealed a main effect of Garden-path for all LMs (albert-base: *estimate* = 1.77, *SE* = 0.29, *t* = 6.09, *p* < 0.001; albert-large: *estimate* = 1.66, *SE* = 0.29, *t* = 5.64, *p* < 0.001; albert-xlarge: *estimate* = 1.80, *SE* = 0.37, *t* = 4.85, *p* < 0.001; albert-xxlarge: *estimate* = 1.16, *SE* = 0.33, *t* = 3.55, *p* < 0.001; bert-base: *estimate* = 1.91, *SE* = 0.34, *t* = 5.59, *p* < 0.001; bert-large: *estimate* = 1.70, *SE* = 0.34, *t* = 5.04, *p* < 0.001; LSTM: *estimate* = 5.05, *SE* = 0.29, *t* = 17.29, *p* < 0.001) and of Transitivity for all LMs (albert-base: *estimate* = 0.92, *SE* = 0.29, *t* = 3.15, *p* < 0.01; albert-large: *estimate* = 1.28, *SE* = 0.29, *t* = 4.36, *p* < 0.001; albert-xlarge: *estimate* = 1.07, *SE* = 0.37, *t* = 2.88, *p* < 0.01; albert-xxlarge: *estimate* = 1.24, *SE* = 0.33, *t* = 3.80, *p* < 0.001; bert-large: *estimate* = 0.98, *SE* = 0.34, *t* = 2.91, *p* < 0.01), except bert-base and LSTM (marginal effect for both; *p* < 0.1).

Along with these main effects, there was also a significant interaction between Transitivity and Garden-path for all models (albert-base: *estimate* = 1.74, *SE* = 0.58, *t* = 2.99, *p* < 0.01; albert-large: *estimate* = 1.92, *SE* = 0.59, *t* =

3.27, $p < 0.01$; albert-xlarge: *estimate* = 1.73, *SE* = 0.74, $t = 2.33$, $p < 0.05$; albert-xxlarge: *estimate* = 2.17, *SE* = 0.66, $t = 3.32$, $p < 0.01$; bert-large: *estimate* = 1.84, *SE* = 0.67, $t = 2.72$, $p < 0.01$; LSTM: *estimate* = 2.29, *SE* = 0.58, $t = 3.93$, $p < 0.001$) with the exception of bert-base (marginal; $p < 0.1$). All LMs had larger surprisal for garden-path condition than no garden-path conditions, this garden-path effect was greater within transitive conditions for most of the models, as revealed in a significant interaction between Transitivity and Garden-path. This imply that LMs are able to utilize comma as an indicator of a clause boundary and sensitive to the linguistic cue of transitivity.

A main effect of Length was found only for LSTM (*estimate* = -0.97, *SE* = 0.29, $t = -3.31$, $p < 0.01$). As in Plausibility test, the surprisal was larger when some intervening words were added. Meanwhile, no interaction between Length and other factors was identified in all LMs.

All models show a generally similar pattern, but LSTM seems to have a relatively large garden-path effect in intransitive conditions compared to Transformers. However, this difference was not statistically significant ($p > 0$), and this difference may be unreliable because the baseline of surprisal may differ across architectures. As for Parameters, there was a main effect of Parameters (*estimate* = -0.004, *SE* = 0.001, $t = -4.53$, $p < 0.001$), which indicates that the larger the number of parameters, the smaller the surprisal regardless of the conditions (Figure 5).

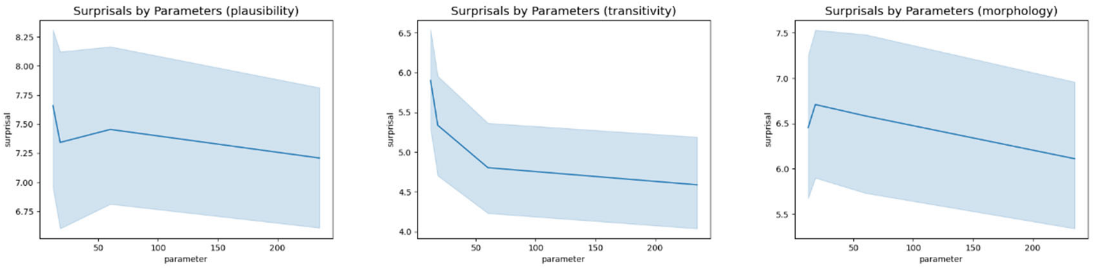


Figure 5. Mean Garden-path Effects by Cue (Plausibility, Transitivity, Morphology) and Length across LMs

3.3 Morphology

For the third test, 28 items such as Example (3) were adapted from Futrell et al. (2019). (3a) and (3b) are reduced relative clauses which are relative clauses sentences with no explicit relative pronoun. Sentences containing a reduced relative clause can be temporarily ambiguous, when the form of the past participle is the same as the past tense verb in English as in (3a). Comprehenders might misinterpret the past participle in the relative clause, *brought*, as the main verb following the subject, *the woman*. This misinterpretation, or temporary ambiguity, would be resolved when they encounter the main verb, *tripped*, which will increase RT/surprisal. On the other hand, (3b) is also a reduced relative clause sentence, but not ambiguous, since the past participle used in this sentence morphologically differs from the past tense verb. In this condition, it is predicted that if LMs are sensitive enough to process the morphological cue, the garden-path effect at the disambiguating verb will be smaller than that in the morphologically ambiguous conditions. The longer version was also created by adding some intervening words to see how it affects LMs' processing.

(3) a. The woman brought the sandwich from the kitchen (with a new oven) tripped on the carpet. [Ambiguous, Garden-path]
b. The woman given the sandwich from the kitchen (with a new oven) tripped on the carpet. [Unambiguous, Garden-path]
c. The woman who was brought the sandwich from the kitchen (with a new oven) tripped on the carpet. [Ambiguous, No garden-path]
d. The woman who was given the sandwich from the kitchen (with a new oven) tripped on the carpet. [Unambiguous, No garden-path]

Figure 6 shows the average garden-path effects at the disambiguating word for both (morphologically) ambiguous and unambiguous conditions. A similar pattern is revealed across all models where both conditions show the garden-path effects and the size of the effect is seemingly larger in the ambiguous conditions.

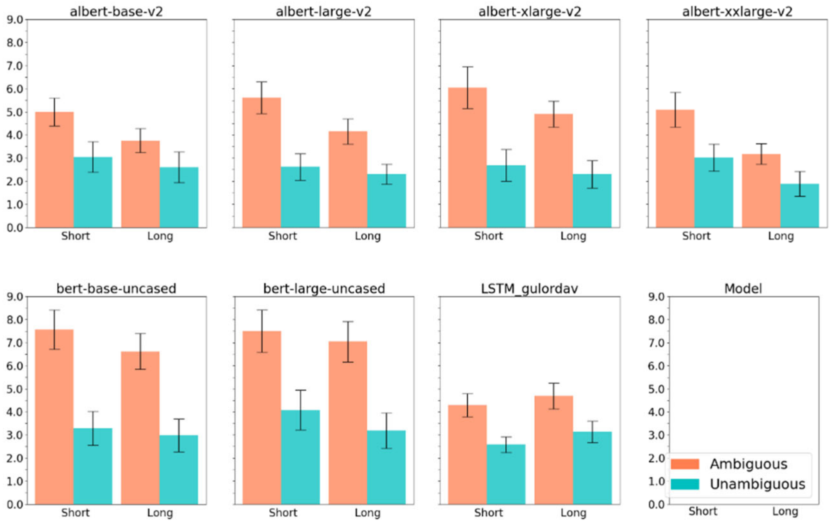


Figure 6. Mean Garden-path Effects by Ambiguity (Morphology) and Length across LMs

There was a main effect of Garden-path for all LMs (albert-base: *estimate* = 3.49, *SE* = 0.35, *t* = 9.86, *p* < 0.001; albert-large: *estimate* = 3.70, *SE* = 0.34, *t* = 10.82, *p* < 0.001; albert-xlarge: *estimate* = 4.10, *SE* = 0.39, *t* = 10.51, *p* < 0.001; albert-xxlarge: *estimate* = 3.11, *SE* = 0.40, *t* = 7.87, *p* < 0.001; bert-base: *estimate* = 5.14, *SE* = 0.42, *t* = 12.30, *p* < 0.001; bert-large: *estimate* = 5.16, *SE* = 0.44, *t* = 11.67, *p* < 0.001; LSTM: *estimate* = 3.80, *SE* = 0.29, *t* = 13.25, *p* < 0.001) and of Morphology for all LMs (albert-large: *estimate* = -1.19, *SE* = 0.34, *t* = -3.49, *p* < 0.001; albert-xlarge: *estimate* = -1.43, *SE* = 0.39, *t* = -3.67, *p* < 0.001; albert-xxlarge: *estimate* = -1.18, *SE* = 0.40, *t* = -2.98, *p* < 0.01; bert-base: *estimate* = -1.94, *SE* = 0.42, *t* = -4.65, *p* < 0.001; bert-large: *estimate* = -1.81, *SE* = 0.44, *t* = -4.09, *p* < 0.001; LSTM: *estimate* = -0.89, *SE* = 0.29, *t* = -3.10, *p* < 0.01) except for albert-base (marginal; *p* < 0.1).

A significant interaction between Morphology and Garden-path was also found for all models (albert-base: *estimate* = 1.51, *SE* = 0.71, *t* = 2.13, *p* < 0.05; albert-large: *estimate* = 2.47, *SE* = 0.68, *t* = 3.60, *p* < 0.001; albert-xlarge: *estimate* = 3.32, *SE* = 0.78, *t* = 4.24, *p* < 0.001; albert-xxlarge: *estimate* = 1.58, *SE* = 0.79, *t* = 2.00, *p* < 0.05; bert-base: *estimate* = 3.64, *SE* = 0.83, *t* = 4.36, *p* < 0.001; bert-large: *estimate* = 3.57, *SE* = 0.88, *t* = 4.04, *p* < 0.001; LSTM: *estimate* = 1.53, *SE* = 0.57, *t*= 2.66, *p* < 0.01). All models displayed greater surprisal in the garden-path condition compared to the no garden-path condition, and this difference was larger in the ambiguous morphology condition. These findings indicate that the LMs have more difficulty processing reduced relative clause structures as with humans and are able to use morphological cues.

A main effect of Length (Figure 7) was identified in two models, albert-base (*estimate* = 0.77, *SE* = 0.35, t = 2.19, $p < 0.05$) and albert-large (*estimate* = 0.76, *SE* = 0.34, t = 2.23, $p < 0.05$). Their surprisal was significantly larger in short conditions than in long conditions, which was the opposite trend of Length effect that the LSTM had in the plausibility and transitivity tests. In those analysis, LSTM showed larger surprisal in the long sentence conditions. In addition, there was a significant interaction between Length and Garden-path for albert-xxlarge (*estimate* = -1.79, *SE* = 0.79, t = -2.26, $p < 0.05$), where the garden-path effect was reduced within longer sentence conditions.

No significant difference was found across the architectures. Nonetheless, there were several noteworthy findings, one of which was that the direction of interaction between Length, Garden-path and Architecture was opposite to that of plausibility (Figure 3). In morphology test, the garden-path effect was smaller in long sentence conditions for Transformers, while larger for LSTM (marginal; $p < 0.1$). However, the reverse was true in the plausibility test. Another finding to note was that the size of the garden-path effect in LSTM was relatively decreased and that of the Transformer (especially BERT; marginal interaction between Morphology, Garden-path and Architecture; $p < 0.1$) was increased compared to the other two tests (Figure 3). Mean surprisal itself was always significantly larger in LSTM in all tests, and thus the magnitude of the garden-path effect was also significantly larger.[3] However, the morphology test showed a reduced garden-path effect for LSTM compared to BERT.

As for Parameter, there was no significant effect or interaction (marginal effect of Parameter; $p < 0.1$).

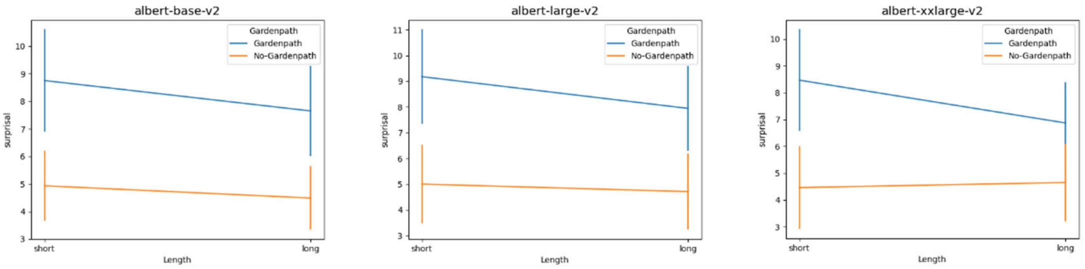


Figure 7. Mean Surprisal by Garden-path and Length for Albert-base (Left), Albert-large (Center) and Albert-xxlarge (Right)

4. General Discussion and Conclusion

This study examined the incremental syntactic processing of neural language models, comparing Transformer models, BERT and ALBERT, with LSTM through a targeted evaluation approach. The first research question was to see whether LMs have incremental syntactic representation comparable with human language processing. The results showed that both Transformers and LSTM performed human-like syntactic processing since they detected temporal ungrammaticality induced by garden-path structures and were capable of using various linguistic cues such as plausibility, transitivity and morphology. Direct comparisons to determine whether it is superior or inferior

[3] The statistics are not reported for the reasons mentioned in the statistical analysis.

to human processing are not possible, due to differences between the dependent measures used in human experiments and this study. However, it can be inferred that at least LMs have comparable syntactic processing with human processors in that the patterns of surprisal across the conditions were similar to those of human reading times.

The second research question was to find out what linguistic cues influence the LMs' syntactic processing. The findings demonstrated that, in transitivity and morphology tests, all LMs generally had syntactic representation influenced by those linguistic cues, despite some minor differences among the models. In each test, the garden-path effect was significantly reduced in intransitive and unambiguous morphology conditions, which are more difficult to induce garden-path. However, only a few models were sensitive to plausibility, showing a significant interaction between Plausibility and Garden-path. The other models than albert-xxlarge and bert-large did not reach the significant level. For this relatively less sensitivity to plausibility, two conflicting interpretations can be discussed. The first is that LMs prioritize syntactic processing. In the transitivity test, which has the same sentence structure with the plausibility test, it is not syntactically possible for an intransitive verb to take an object, whereas syntactically possible for an implausible verb to have an object but it is merely difficult to be semantically connected with the subsequent noun. That is, misinterpretation is syntactically probable even in implausible conditions. In fact, several human psycholinguistic studies (Roberts and Felser 2011) demonstrated that the plausibility effect did not occur depending on the sentence structure. From the classical view of Frazier (1987) on sentence processing, the initial state in sentence processing is based solely on a syntactic ground, unaffected by other semantic or pragmatic factors. From this perspective, it may be considered that LMs did not fail to employ plausibility cue, but instead prioritized syntactic analysis despite being able to do so. Another interpretation is that LMs are relatively poor at utilizing information about semantic connections between words. Among the three linguistic cues, plausibility is likely to be the most relevant to semantic properties. For transitivity and morphology, information on the morpho-syntactical properties of the verb is important, while for plausibility, knowledge about the semantic relationship between the verb and the following noun is required. It is to some extent consistent with the previous study (Tenney et al. 2019) where contextualized embeddings such as BERT showed greater improvement on syntactic tasks than on semantic tasks. It is difficult to conclude which of the two interpretations is more likely from the current findings. Nevertheless, it can be assumed that it is related to the capacity to exploit sematic information in that only models with greater performance than other sub-models in the majority of tests, such as albert-xxlarge and bert-large, showed a significant difference in garden-path effect by plausibility.

The third question was how the syntactic representations of more recent models such as ALBERT and BERT are different from more traditional LSTM models in terms of human-like syntactic processing. Overall, no statistically significant pattern difference was found between the three architectures. All LMs exhibited garden-path effects in all tests and demonstrate the ability to utilize linguistic cues except plausibility. One difference between architectures was the sensitivity to plausibility. Although not statistically significant, there was virtually no difference in the garden-path effect between the plausible and implausible conditions in LSTM, whereas the sub-models of Transformers showed at least a numerically less garden-path effect in the implausible conditions, and a significant interaction between Garden-path and Plausibility was found in two Transformers, albert-xxlarge and bert-large. Despite some conflicting results, numerous psycholinguistic studies show that human syntactic processing is rapidly affected by discourse and semantic context (Pickering and Traxler 1998). In this respect, Transformers may demonstrate a more human-like syntactic representation compared to LSTM, at least in plausibility test.

The present study also explored how intervening words affect the syntactic performance of language models and what differences exist between the models. For Transformer models, the results did not reveal an effect of

Length in plausibility and transitivity manipulation except that albert-large showed greater surprisal in long sentence conditions. On the other hand, in the case of LSTM, a length effect was observed both in plausibility and transitivity manipulation. Moreover, in the plausibility test, the garden-path effect for LSTM was reduced in longer sentences, supported by a significant interaction between Plausibility and Length, which is the opposite of what is often predicted for human processing. In human processing, the garden-path effect increases with intervening words because the longer the subsequent noun is misinterpreted as an object, the more difficult it is to recover from the misinterpretation and reanalyze it to the correct meaning (Ferreira and Henderson 1991). This suggests that the garden-path effect is more pronounced in extended sentences only if information on prior words is preserved. In the LSTM model, it is known that as the sentence length increases, the influence of the previous words is reduced (loss of the weight of the previous words). Thus, intervening words reduced the predictive power for upcoming words in general, which may have produced the length effect of the LSTM. In addition, the garden-path effect was also reduced because information about the preceding word was lost. In contrast, Transformers did not show a difference in surprisal by Length except one sub-model in plausibility test. Furthermore, there was a significant interaction among Length, Garden-path and Architecture, which suggested that in Transformers, the garden-path effect increased with the addition of words in contrast to LSTM. This might indicate the attention technique applied to Transformer has improved the weight loss of the previous words.

However, in the morphology manipulation structure, there was a length effect only for some of Transformers, but not in LSTM. Unexpectedly, average surprisal in short sentences was rather larger compared to the one in long sentences. Smaller surprisal might indicate better prediction, because surprisal is inverse logarithm of the probability. In fact, more words mean more information, so there is a possibility that predictive power can increase when several words are added, if information can be appropriately used. Transformers may have benefited from the addition of words because the loss of information on previous words is relatively small compared to LSTM. However, increasing the number of words did not always improve the predictive power of Transformers. In the plausibility test, surprisal increased in long sentences, and there was no difference in transitivity. The effect may vary depending on what type of words or how many words are added. In this study, it is difficult to clearly interpret this due to the lack of control over the type or number of words to be added. However, several possibilities can be discussed. Above all, it does not seem that simply having a large number of words provides a predictive advantage. Although the distance between the preceding noun and the disambiguating word was larger in morphological manipulation than in the others, there was no effect of length in transitivity, where the number of words similar to plausibility was added. One thing to note is that this reverse length effect appeared along with an interaction between Length and Garden-path. As shown in Figure 7, there was a smaller garden-path effect in the long sentence condition, which seems to be driven by the decrease of surprisal in the garden-path condition. The additional words might provide some information to resolve earlier misinterpretation, aiding in the resolution of the garden-path effect, which might in turn lead to a decrease of the overall surprisal in long conditions. As mentioned in the case of LSTM, a smaller garden-path effect in the long conditions may indicate loss of information on previous words. However, the smaller effect in LSTM was mainly due to the surprisal increase in the no garden-path condition within long sentences, along with overall increase of surprisal (Figure 2). In the case of Transformers, it was derived from smaller surprisal within garden-path conditions in conjunction with overall surprisal reduction. Therefore, it is presumed that the lessened garden-path effect was caused by faster resolution of difficulty in misinterpretation as a result of the additional information, rather than due to the loss of information on the previous words. Similarly, in psycholinguistic research, it was anticipated that reanalysis or revision of misinterpretation can be often facilitated in several conditions and performed even before the disambiguation words (Pickering and Traxler 1998)

The fourth research question was whether the different number of parameters of pre-trained models affects their syntactic processing. Overall, a clear difference in syntactic processing according to the different number of parameters was not found. However, at least it can be concluded that the larger number of parameters does not contribute to less human-like syntactic processing as suggested in the previous study (Wilcox, Levy and Futrell 2019). Rather, albert-xxlarge and bert-large, which has the largest number of parameters among the models, showed more human-like syntactic representation in plausibility test. Moreover, since the surprisal tended to be smaller as the parameters increased, more parameters may improve the predictive power of models. Nevertheless, it is also not that the more parameters, the better, since all Transformers revealed similar patterns in terms of human-like syntactic representation. Thus, it seems that the relationship between parameters and syntactic performance is not a simple, positive or negative linear relation, but more complicated, requiring further investigation.

To conclude, the present study investigated the incremental syntactic processing of neural language models such as BERT, ALBERT, and LSTM, through a targeted evaluation approach. Both Transformers and LSTM, when processing garden-path structures, were capable of employing language cues such as clausal boundary, relative clause, plausibility, transitivity, and morphology, which were comparable to human language processing. Although all models showed in general a similar pattern of syntactic representations, there were also notable discrepancies between Transformers and LSTM. Some of Transformers appear to be more sensitive to plausibility than LSTM and utilize the information of previous words more effectively for language processing. Meanwhile, the number of parameters did not seem to have great influence on LMs' syntactic processing. However, at least more parameters did not adversely affect their processing, since the increase in parameters led to the decrease in surprisal and the model with the most parameters was the most sensitive to semantic information such as plausibility.

Through these findings, this study aimed to contribute to a deeper understanding of the syntactic processing of deep learning neural language models as well as human language processing. Currently, neural language models have been developing rapidly, and often show superior performance over humans in several language performance tests. However, there is a paucity of knowledge on how such abilities were acquired; hence they remain as a black box. Moreover, pre-trained LMs learned a syntactic structure and had a high level of semantic representation only through a significant amount of unprocessed, unstructured raw data. This remarkable ability to learn had linguists rethink the fundamental mechanism underlying language acquisition. Using psycholinguistic research methodologies to evaluate language models is an attempt to evaluate the linguistic representations of language models at a higher level and compare them with those of humans. By examining how LMs process sentences that are challenging even for people, it is possible to explore if they are developing in a manner analogous to human language processing or if they are improving performance in a different way from humans. While this will reveal the deficiencies of LMs that need to be further addressed, at the same time, it will bring new insight about human language processing. This study tried to develop a discussion from this point of view. Although it is difficult to draw clear conclusions from the current findings alone, the tentative conclusion is that LMs are developing syntactic representations comparable to human syntactic representations, and that more recent models with greater performance are more similar to human syntactic processing.

In future research, the significance of this study should be developed by addressing the limitations shown in these studies. This research had several limitations. One is that the comparison between LMs and human language processing was indirect. Although the majority of the sentences employed in this study were adapted from psycholinguistic experiments, they were not identical with the previous materials. It is adequate for a rough comparison with human language processing, but it is not a direct comparison. A more accurate comparison would be possible if the experiment was conducted with the same sentence both on human participants and RTs and surprisal were compared. Another limitation is the lack of statistical comparisons between model architectures. Due to the varied vocabulary of the models, it was not feasible to directly compare surprisal. It is necessary to adjust the number of the words for more objective model comparisons. Finally, regulation over the attributes of linguistic cues or additional words is needed. In this study, differences according to sentence length were identified across the architectures, but it was difficult to interpret those differences because the intervening words were syntactically and semantically heterogeneous. Fine-grained control over the sentence may facilitate the interpretation of the findings. By overcoming these limitations, the evaluation of the language model through the research method of psycholinguistics will provide a valuable perspective on the internal structure of neural language models.

References

Adams, B. C., C. Clifton and D. C. Mitchell. 1998. Lexical guidance in sentence processing? *Psychonomic Bulletin & Review* 5(2), 265-270.

Baayen, R. H., D. J. Davidson and D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4), 390-412.

Bever, T. G. 1970. Cognitive basis for linguistic structures. In J. R. Hayes, ed., *Cognition and the Development of Language*, 279-362. New York: Wiley.

Devlin, J., M. W. Chang, K. Lee and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805*

Demberg, V. and F. Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109(2), 193-210.

Ettinger, A. 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.

Frazier, L. 1987. Sentence processing: A tutorial review. In M. Coltheart, ed., *Attention and Performance 12: The Psychology of Reading*, 559-586. Lawrence Erlbaum Associates, Inc.

Frazier, L. and K. Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology* 14(2), 178-210.

Ferreira, F. and J. M. Henderson. 1991. Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language* 30(6), 725-745.

Futrell, R., E. Wilcox, T. Morita and R. Levy. 2018. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency*. arXiv preprint arXiv:1809.01329.*

Futrell, R., E. Wilcox, T. Morita, P. Qian, M. Ballesteros and R. Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260.*

Goldberg, Y. 2019. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287.*

Goodkind, A. and K. Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. *In Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics* (CMCL 2018), 10-18.

Gulordava, K., P. Bojanowski, E. Grave, T. Linzen and M. Baroni. 2018. Colorless green recurrent networks dream hierarchically*. arXiv preprint arXiv:1803.11138.*

Hu, J., J. Gauthier, P. Qian, E. Wilcox and R. P. Levy. 2020. A systematic assessment of syntactic generalization in neural language models. arXiv preprint arXiv:2005.03692.

Jegerski, J. 2013. Self-paced reading. In J. Jegerski and B. VanPatten, eds., *Research Methods in Second Language Psycholinguistics*, 36-65. Routledge.

Just, M. A. and P. A. Carpenter. 1980. A theory of reading: from eye fixations to comprehension. *Psychological Review* 87(4), 329-354.

Kuncoro, A., L. Kong, D. Fried, D. Yogatama, L. Rimell, C. Dyer and P. Blunsom. 2020. *Syntactic Structure Distillation Pretraining for Bidirectional Encoders*. *arXiv preprint arXiv:2005.13482.*

Kuznetsova, A., P. B. Brockhoff and R. H. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82, 1-26.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942.*

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

Linzen, T., E. Dupoux and Y. Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4, 521-535.

Marvin, R. and T. Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031.*

Pickering, M. J. and M. J. Traxler. 1998. Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24(4), 940.

Roberts, L. and C. Felser. 2011. Plausibility and recovery from garden paths in second language sentence processing. *Applied Psycholinguistics* 32(2), 299-331.

Smith, N. J. and R. Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128(3), 302-319.

Staub, A. 2007. The parser doesn't ignore intransitivity, after all. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33(3), 550.

Tenney, I., P. Xia, B. Chen, A. Wang, A. Poliak, R. T. McCoy ... and E. Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905. 06316*.

Trueswell, J. C., M. K. Tanenhaus and S. M. Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language* 33(3), 285-318.

van Gompel, R. P. and M. J. Pickering. 2001. Lexical guidance in sentence processing: A note on Adams, Clifton, and Mitchell (1998). *Psychonomic Bulletin & Review* 8(4), 851-857.

van Schijndel, M. and T. Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. *CogSci*.

van Schijndel, M., A. Mueller and T. Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. *arXiv preprint arXiv:1909.00111.*

Wilcox, E., R. Levy, T. Morita and R. Futrell. 2018. *What do RNN Language Models Learn about Filler-Gap Dependencies? arXiv preprint arXiv:1809.00042.*

Wilcox, E., R. Levy and R. Futrell. 2019. Hierarchical representation in neural language models: Suppression and recovery of expectations. *arXiv preprint arXiv:1906.04068.*

Examples in: English
Applicable Languages: English
Applicable Level: Tertiary