Plausibility and Well-formedness Acceptability Test on Deep Neural Nativeness Classification

Kwonsik Park
Korea University
Department of Linguistics
Oneiric66@korea.ac.kr

Sanghoun Song
Korea University
Department of Linguistics
sanghoun@korea.ac.kr

Abstract

The present work compares performance among several deep learning models, of which task is to clasify nativeness of English sentences. The current study constructs 4 models, each using different deep learning networks: RNN, LSTM, BERT and XLNet. We use 3 test suites to evaluate the four models: (i) 8 test sets composed of 4 native- and non-native-written data, (ii) a supplemented version of well-formedness and plausibility test set consisting of 120 sentences from Park et al. (2020), and (iii) a test set of 196 sentences consisting of 11 types (27 subtypes) for grammaticality judgment test (DeKeyser, 2000). The results show that the more up-to-date models, BERT and XLNet outdo relatively out-of-date models, RNN and LSTM. The latest model among the 4 models is XLNet, but it does not outperform BERT in every aspect. Presuming that the ways deep learning learns language are, to some extent, similar to the strategies of L2 learners, the current work trains the models with data consisting of native and learner English sentences to compare nativeness judgments between deep learning models and humans for investigating if it is the case. This paper concludes that there are few learnability problems shared by the two agents.

1 Introduction

Deep learning is no longer an unfamiliar word to NLP researchers. It has already been used in various tasks in NLP such as the ways to resolve long-distance filler-gap dependencies (Da costa & Chaves, 2020; Chaves, 2020; Wilcox et al, 2019), number agreement (Linzen & Leonard, 2018), reflexive anaphora (Goldberg, 2019), to name a few, and shown lots of striking results. However, there are very few works attempting to train deep learning with learner corpora.

It seems that the ways deep learning learns language are similar to L2 learners' language learning strategies in that both artificial and natural intelligence generalize data, extract meaningful features, solve problems, check errors, modify what has been their knowledge and memorize what they have learned from this procedure. This is in line with the Fundamental Difference Hypothesis (Bley-Vroman, 1988). The hypothesis argues that adult learners learn language with analytical, problem-solving mechanisms. In addition, they are also alike in that their language learning is limited by the poverty-of-the-stimulus, i.e., they depend mainly on input data from the outside and cannot learn a language completely without rich enough data whereas children can.

To check if it is the case, this work makes four language models built up with four different artificial neural network, Recurrent Neural Network (RNN, Mikolov et al., 2010), Long-Short Term Memory RNN (LSTM, Sundermeyer et al., 2012), Bidirectional Encoder Representations from Transformers (BERT, Devlin et al., 2018), and XLNet (Yang et al., 2019). The task of the models is nativeness classification. Prior works such as Warsdadt et al. (2019) construct their classification models trained with L1 data labeled with a binary grammaticality value. To some extent, their models learn unacceptability of sentences because they are trained with ungrammatical sentences, but we

cannot say they learn *learners*. Instead, we train our models with L1 and L2 data to learn L2 learners. Each of the four models is then evaluated by classifying nativeness of every sentence in 3 test suites to examine the deep learning models in various angles.

Type	Test Data	Number of Sentences	Average Sentence Length
Native (diplomatic, news)	The English Gigaword	3,000	24.83
	The Europarl Corpus	3,000	23.32
Native (novel)	The Speckled Band	572	17.16
	The Little Prince	1,835	9.12
Non-Native (elaborated)	The Three English Speeches of President Kim Dae-Jung	484	17.51
	The Tanaka Corpus	3,000	7.79
Non-Native (not elaborated)	INUMLC (spoken)	697	4.55
	INUMLC (written)	613	14.52

Table 1: The 8 test sets

2 Nativeness Classification

The task assigned to the four models is to identify nativeness of sentences, i.e., to predict whether a sentence is written by a native or non-native speaker. Pawley and Syder (1983) observe that nativelike sentences be 'institutionalized' and 'lexicalized' patterns. They also note that two essentials are required for an expression to be a 'native selection'; not only should it be grammatically well-formed but sounds plausible. An ill-formed sentence refers to the one that has syntactic violations, and an implausible sentence is the one that is syntactically well-formed but sounds awkward to native speakers, e.g., "*I wish to be wedded to you."* (Pawley and Syder, 1983) does not sound natural despite its well-formedness.

3 Model Construction

The present work constructs deep learning models which are trained with native and non-native data to predict nativeness of sentences using four types of deep learning networks: RNN, LSTM, BERT and XLNet. As mentioned above, the task is binary classification of nativeness. The models are designed to output '1' when they predict a sentence as native one, and '0' when predict it as a non-native one. RNN and LSTM models were trained for 10 epochs, and among the 10 epochs, the model with the highest validation accuracy was the highest was used (BERT and XLNet for 4 epochs).

3.1 Data

The entire data are composed of 651,665 sentences, which consist of two parts, training and validation data, and the test data consists of three test suites.

Training and Validation Data

Training data is made up of native- and learner-written sentences, the total size of which is 586,501 sentences (7,852,306 words). The native data used in this paper is extracted from the Corpus of Contemporary American English (COCA). The learner data is excerpted from Yonsei English Learner Corpus (YELC) and Gacheon Learner Corpus (GLC), both of which were made by undergraduate students in Yonsei University and Gacheon University in Korea, respectively. Validation data is one-tenth of the entire data (65,164 sentences), which is used for evaluating classification accuracy of the deep learning models.

Test data

Test data consists of three test suites: (i) 8 test sets, (ii) a supplemented version of well-formedness and plausibility test items consisting of 120 sentences from Park et al. (2020) and (iii) grammaticality judgment test items composed of 196 test items from DeKeyser (2000).

The 8 test sets are made up of 4 native- and 4 non-native-written test sets. As shown in Figure 1, Each of The English Gigaword and The Europarl

Corpus is a highly elaborated version of native randomly excerpted from original data, the former being news and the latter being diplomatic sentences. The Speckled Band and The Little Prince are novel data. English Speeches of President Kim Dae-Jung (hereafter, KDJ) was revised by Korean diplomatic experts, so it has no syntactic violation. The Tanaka Corpus is composed of 3,000 sentences randomly extracted from original data. This was edited by researchers in the process of corpus construction, so this corpus also does not have any syntactic violation. In contrast, both the written version of Incheon National University Multilingual Learners Corpus (hereafter, INUMLC (written)) and the spoken version of Incheon National University Multilingual Learners Corpus (hereafter, INUMLC (spoken)) have lots of syntactic errors in them.

The second test suite English test items from Park et al. (2020), which originally were designed to compare well-formedness and plausibility judgments of native English subjects to those of a deep learning model (the model used in the paper was RNN). The test items of the article consist of controlled and filler items. We use only the controlled items because the filler set was made for the language experiment. The controlled items are composed of 60 well-formedness and 50 plausibility test items, so we add 10 plausibility test sentences to balance between them. Consequently, the test suite consists of 120 sentences.

The last test suite is test items from Dekeyser (2000), which are made up of 196 well-formedness test items categorized into 11 types (27 subtypes). The test items are originally made for examining English grammaticality judgments of immigrants living in the United States (Dekeyser, 2000). The second and third test suites are made up of pairwise sentences and each is labeled in a binary way: ill-formed sentence is labeled '0' and well-formed is '1'. The second suite is not composed of minimal pairs while the third one is. Test items from Dekeyser (2000) are used to investigate whether each model shows similar judgment patterns to those of L2 learners (immigrants).

3.2 Four types of networks

The present study constructs four models, each of which is made from the four different deep learning networks.

Firstly, RNN is a type of artificial neural network in which information from a previous step data. Each of them is composed of 3,000 sentences is updated in the current step but has a limitation of gradient vanishing, which refers to the problem that the longer steps information is carried over, the more information the model loses.

Secondly, LSTM is an elaborated version of RNN. It resolves the problem of gradient vanishing to some extent by updating information from previous steps selectively; important one is memorized and not important one is discarded.

Thirdly, BERT is a relatively up-to-date neural network that 'is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers' (Devlin et al., 2018), which means it can entail bidirectional contextual information on each word by using a special noise token, [MASK], which should be predicted by the model, resulting in representing rich information. BERT is so powerful that it outdoes performance of many previous NLP models, but it has a limitation that 'BERT assumes the predicted tokens are independent of each other given the unmasked tokens, which is oversimplified as high-order, long-range dependency is prevalent in natural language' (Yang et al., 2019).

Lastly, to solve this problem, Yang et al. elaborated BERT to build up the XLNet network, which is capable of doing both autoregressive and autoencoding methods by '[maximizing] the expected log likelihood of a sequence w.r.t. all possible permutations of the factorization order'.

The present work constructs four models using those four kinds of deep learning networks and investigates which model is better to detect nativeness of English sentences.

4 Results

The 4 models are evaluated on 3 test suites by classifying nativeness of every sentence in each test suite.

4.1 Test Suite I

Table 2 shows the results of evaluation on the 8 test sets. Numbers in the table are proportions of sentences which are predicted as native sentences in each test set. Left 4 test sets are native data, so a higher score means better performance whereas for 4 test sets in the right side, a lower score means better performance.

	Native Data				Non-native Data			
	The English Gigaword	**The Europarl Corpus**	**The Speckled Band**	**The Little Prince**	**KDJ**	**The Tanaka Corpus**	**INUMLC (spoken)**	**INUMLC (written)**
RNN	97.9	95.7	95.8	91	95.8	83.4	82	44.6
LSTM	95.8	92.4	88.1	76.8	86.7	66.4	76.7	30.8
BERT	95.5	90.1	92.6	80.1	**73.5**	**53.2**	**57.8**	12
XLNet	97.1	94.6	96.1	85.8	76.4	59.8	64.4	**7.6**

Table 2: Nativeness judgment results on the 8 test sets
(proportions of sentences which are predicted as native sentences)

The 4 models predict the nativeness of native test sets well in general except for The Little Prince. The low accuracy on The Little Prince is probably caused by two reasons: the first one is its sentence length is relatively short (see Table 1, the average sentence length is 9.12), which means each sentence in the novel has relatively few clues for the models to make use of to predict nativeness, and the other one is that it has somewhat learner-like sentences which do not contain difficult words or complex clauses. For example, (1a) and (1b) are the ones that every model predicts as non-native sentences. We do not say, of course, sentence length is not a necessary condition of nativeness. But the short sentence length must have influenced the models' judgments.

(1) a. It is unnecessary.
b. This is a ram.

With respect to 4 non-native test sets, there are remarkable differences among test sets. Regarding KDJ and The Tanaka Corpus, the 4 models give relatively high scores to them although they are non-native data. This is probably because, as mentioned before, they are manually edited learner corpora that have no syntactic violation. The reason that KDJ is given higher scores than The Tanaka Corpus could be explained in terms of plausibility; the editing on learner sentences in The Tanaka Corpus was focused only on their syntactic well-formedness, whereas KDJ was sophisticatedly edited on both grammaticality and its content, i.e., its well-formedness and plausibility. Furthermore, The Tanaka Corpus has more learner-like sentences than KDJ because KDJ is such a diplomatic document that it has few learner-like sentences that L2 learners use in everyday conversation. (2a) and (2b) are the ones in The Tanaka Corpus that every model classifies as non-native sentences, partially because of its awkwardness or learner-likeness.

(2) a. My father is proud of my being handsome.
b. My uncle gave me a book.

(2a) is syntactically well-formed but not made in a frequently used pattern. (2b) has no syntactic violation, too, but it is the one that usually appears in learner textbooks. To sum up, KDJ gets higher score as it was elaborated in terms of plausibility as well as well-formedness of its sentences. Nevertheless, the nativeness of KDJ is not fully satisfied; it does not get the score as high as that of native data from models except for RNN, which indicates, however elaborated a text is, it's almost impossible for non-natives to reach the level of native speakers (Park et al. 2019). The scores of 4 models show that BERT and XLNet can detect the nativeness of KDJ and The Tanaka Corpus better than RNN and LSTM.

This better performance of BERT and XLNet is clearly found in the results of INUMLC (written). INUMLC (written) is the most learner-like text which has lots of ill-formed and implausible sentences. These are instantiated in (3a) ~ (3c).

(3) a. *I worried and tired because a lot of people.
b. *I think I could wrote various articles.
c. In my life I like to read books

(3a) has syntactic violations of omitting *be*-verb and using *because* instead of *because of*. (3b) also has a syntactic violation of using a past verb *wrote* after a modal verb *could*. Although (3c) is syntactically well-formed, it is predicted as a non-

	RNN	LSTM	BERT	XLNet
Well-formedness Judgment items	56.6%	56.6%	70.0%	**73.3%**
Plausibility Judgment Items	55.0%	60.0%	**63.3%**	56.6%
Well-formedness + Plausibility	55.8%	58.3%	**66.6%**	65.0%

Table 3: Nativeness judgment results on test items from Test Suite II

native sentence by all the models, partially because *in my life* is not plausible; not only does the expression not go well with the context, but it also usually occurs in the last position of a clause. In this sense, the performance of the four models is reflected the most in the scores of INUMLC (written) because models' judgments on this pure learner data show how correctly they can discriminate non-native sentences from native ones. The scores of BERT and XLNet show a drastic decrease from those of RNN and LSTM, and the latest model, XLNet, gives the lowest score to INUMLC (written), indicating it has the highest performance on the test set.

INUMLC (spoken), on the other hand, gains scores from the models that do not accord with our intuition; although syntactic violations and awkward expressions are more common in spoken data, the scores are quite high, which means all the four models cannot correctly classify the nativeness of the test set. An explanation is that the average sentence length of INUMLC (written), 4.55 words, is too short for the models to detect clues to use for sentences classification. This phenomenon is also found in Park et al. (2019), where an RNN model is used for nativeness classification. (4a) ~ (4c) are the examples of short sentences that all models classify as native sentences.

(4) a. Uh, okay.
b. Oh, yeah.
c. Okay.

Whether sentences above are made by natives or non-natives is probably hard to predict even for humans. We are not arguing, of course, that short sentences are the only factor that causes the models to incorrectly judge the test set; it is just one variable that influences the predictions.

In sum, the results on the 8 test sets indicate that XLNet and BERT have better performance than RNN and LSTM, and all the models seem to consider plausibility (i.e., no awkwardness) as well as well-formedness (i.e., no syntactic violation) of sentences. This demonstrates that deep learning can learn syntactic and, by extension, beyond syntactic information from native and learner data. This is reasonable because the four models learn information of a word by considering the words surrounding it, which is like the way Firth (1961) put forward: 'You shall know a word by the company it keeps'. This proposes the possibility of investigating nativeness that has been tricky and hard to prove.

One limitation of the analyses on Test Suite I is we cannot confirm why the models do not exactly classify INUMLC (spoken) and The Little Prince, both of which have relatively short sentences. To investigate short sentence length really confuses deep learning's judgments, in the future research, it is needed to exclude short sentences by establishing a certain threshold of length and compare results to those of this experiment.

4.2 Test Suite II

Table 3 shows the nativeness classification results on a supplemented version of test items from Park et al. (2020). As shown in the table, the accuracy of BERT and XLNet (66.6% and 65%, respectively) is again higher than RNN and LSTM (55.8% and 58.3%, respectively) on the entire test items.

The well-formedness judgment test items consist of 60 sentences that are subcategorized into (i) negative frequency adverb, (ii) the use of *hardly*, (iii) the collocation of the and same, (iv) overpassivization, (v) the use of middle verb, and (iv) be-insertion. Ill-formed sentences of each subcategory are instantiated in (5a) ~ (5f).

(5) a. *You hardly can breathe.
b. *John finds it hardly to talk with strangers.
c. *Mary felt same way about the incident.

		Well-formedness Test Items	Plausibility Test Items
Human Judgements (Park et al., 2020)		55/60 (91.6%)	35/50 (70.0%)
Deep Learning Judgments	RNN	34/60 (56.6%)	26/50 (52.0%)
	LSTM	34/60 (56.6%)	29/50 (58.0%)
	BERT	42/60 (70.0%)	**33/50 (66.0%)**
	XLNet	**44/60 (73.3%)**	30/50 (60.0%)

Table 4: Humans judgments in Park et al. (2020) and deep learning judgments in this paper

d. *A table was appeared.
e. *The articles are translating easily.
f. *Mary is drink water.

XLNet performs better than any other model on the well-formedness test items. If we exclude the well-formedness judgments on adverbs, the accuracy of every model increases: RNN/LSTM: 62.5%, BERT: 80%, XLNet: 85%, which seems to be caused by the difficulty of learning adverb positions; adverbs in English are relatively free in word order, and some adverbs such as negative frequency adverbs are strictly restricted to use while some are not, so deep learning probably feels hard to learn the proper usage of adverbs.

The plausibility judgment items are composed of 60 sentences that are subcategorized into (i) semantic prosody, (ii) semantic preference, (iii) the position of adverbs, (iv) the position of actually, (v) overcomplexity and (vi) collocation of words, each of which consists of 10 sentences. The sixth one is added in the current work to balance the number of plausibility test items with that of well-formedness test items. Implausible sentences of each category are exemplified in (6).

(6) a. A fantastic feast broken out.
b. The company is undergoing customer praise.
c. Also, I hope you're coming to our party tonight.
d. I made my car repaired, actually.
e. That I meet you makes me so happy.
f. Tom ate the pill.

XLNet is not the best but ranks third among the four models, which means XLNet tends to focus mainly on well-formedness of sentences rather than plausibility when predicting nativeness. BERT, On the other hand, gains the highest accuracy among the models, which shows BERT seems to have a more comprehensive view that considers both well-formedness and plausibility of sentences. The results can explain why BERT classifies KDJ and The Tanaka Corpus slightly better than XLNet (see Table 2): KDJ and The Tanaka Corpus are both syntactically well-formed, so from the perspective of XLNet, they deserve higher scores.

Notably, of plausibility test items, the ones for checking the knowledge about overcomplexity are particularly hard for the models to predict them correctly; the implausible items that all four models wrongly classify as plausible are just five sentences, three of which are overcomplex sentences. The three are instantiated in (7)

(7) a. It's the day before Monday.
b. John's becoming Mary's spouse is what he wants.
c. It's one-half of ten dollars.

Every model classifies them as native sentences, but native speakers feel awkward when reading them. An explanation for their wrong prediction is such that deep learning lacks a sense of economy. It is probably true that learners are not capable of making such sentences due to its syntactic complexity, so deep learning is likely to judge such syntactic complexity as a standard of native sentences.

As shown in Table 4 that compares human judgments in Park et al. (2020) to deep learning judgments in this paper, the results show that the accuracy of human judgments overwhelms that of deep learning judgments on well-formedness test items. XLNet is the closest one to native speakers, but there still be a big gap between them. Notably, on the other hand, there is not such a big difference between them on plausibility items; BERT, which is the most sensitive to plausibility as mentioned before, almost reaches the correct rate of humans (the difference of the number of correctly classified sentences is just two items).

	RNN	LSTM	BERT	XLNet
Individual Sentences (196 items)	50.5%	46.4%	54.5%	**56.1%**
Minimal Pairs (98 pairs)	4 pairs	4 pairs	**16 pairs**	14 pairs

Table 5: Nativeness judgment results on test items from DeKeyser (2000)

The participants in the paper are native English speakers. The current study, by extension, compares judgments of English learners to that of deep learning to investigate if there are any shared learnability problems. The next chapter is where this comparison is carried out.

4.3 Test Suite III

The grammaticality judgment test items from DeKeyser (2000) are pairwise minimal pair items consisting of 196 sentences (98 minimal pairs). Categorized into 11 types (27 subtypes), the test set is a highly refined set of items to measure test takers' knowledge in a wide variety of grammatical aspects. The term 'grammaticality' in the article is compatible with 'syntactic well-formedness' in the current paper, that is, this test suite is designed to consider only syntactic well-formedness rather than plausibility of sentences. In this sense, the most critical difference between Test Suite II and Test Suite III is whether they include implausible sentences or not. This test set is chosen for two reasons: (i) to examine deep learning's syntactic knowledge from a more integrated view and (ii) to investigate if there exist common learnability problems that both deep learning and L2 learners experience.

As shown in Table 5, the accuracies of every model are lower than what the models have on the test set from Park et al. (2020), which reveals the limitation that our models haven't learned various syntactic information. (8a) ~ (8k) are examples of 11 types for testing the knowledge of well-formedness.

(8) a. *Last night the old lady die in her sleep. (past tense)
b. *Three boy played on the swings in the park. (plural noun)
c. *John's dog always wait for him at the corner. (third-person singular)
d. *The little boy is speak to a policeman. (present progressive)
e. *Tom is reading book in the bathtub. (determiners)
f. *Peter made out the check but didn't sign. (pronominalization)
g. *The man climbed the ladder up carefully. (particle movement)
h. *George says much too softly. (subcategorization)
i. *Will be Harry blamed for the accident? (yes-no questions)
j. *What Martha is bringing to the party? (wh-question)
k. *The dinner the man burned. (word order)

Suppose the model had failed to learn the grammatical taxonomy of syntax, it would resort to simple heuristics that return probability value of plausibility when judging nativeness of sentences. The test set is largely composed of plausible sentences, so without knowledge of well-formedness, models are likely to classify them as native sentences. This is reflected in the number of sentences that each model predicts as native ones: RNN (159/196), LSTM (125/196), BERT (91/196), and XLNet (134/196). XLNet, which is relatively weak to capture plausibility, predicts the test items as native sentences even more frequently than LSTM (the accuracy of LSTM is also higher than XLNet on plausibility test items of Test Suite II, see Table 2). The results of RNN indicate that it is quite biased to classifying sentences as native ones compared to those of the other models, which is also shown in the 8 test sets (see Table 2); the gap between the highest and the lowest score that RNN gives to the 8 test sets is smaller than any other model. BERT, on the other hand, is the only model that the number of sentences predicted as native ones is lower than half of the test set. This indicates BERT does not have a biased judgement standard compared to the other models.

Minimal sentence pairs that the models classify both correctly are considerably low: RNN, LSTM, BERT and XLNet correctly predict just 4, 4, 16 and 14 pairs out of 98 pairs, respectively. Although the sentences correctly classified by BERT and XLNet are almost 4 times more than those by RNN

and LSTM, BERT and XLNet in the current work are still not good classifiers.

In DeKeyser (2000), the author sorts the test items into three groups into high (difficult), marginal (middle) and low (easy) groups based on the test results from the participants; if the difference of answers to a question among participants is big, the question is distinguished into the difficult group, and if answers to a question from most participant are similar, it is classified into the low group. We investigate if there are any common learnability problems between the subjects and our models; if what has been learned by deep learning and L2 learners are similar, learnability problems of the two agents are also likely to be shared. Presuming minimal pairs that the models predict both correctly are the ones that each model certainly has learned, the current study examines how the pairs are spread among the three levels of difficulty. The number of correctly predicted minimal pairs of 4 models is shown in Table 6.

	RNN	**LSTM**	**BERT**	**XLNet**
High	3/4	2/4	5/16	4/14
Marginal	0/4	0/4	6/16	3/14
Low	1/4	2/4	5/16	7/14

Table 6: Number of correctly classified minimal pairs

It seems that there are few learnability problems that deep learning and learners share because if there are shared problems, the correctly predicted pairs should have been the most in the easy group. From this result, we can assume that the aspects of learning of deep learning and learners are rather different.

5 Discussion and Conclusion

Much of related literature focuses only on syntactic well-formedness. This is partially because plausibility or nativelikeness is hard to define, and numerous contextual factors intervene clear categorization of them. However, for a sentence to be literally *well-formed*, satisfying syntactic rules only is not enough. In this sense, we attempt to examine sentences considering both syntactic well-formedness and plausibility.

The four models in this paper are evaluated from three angles. Firstly, the current work examined whether (and how) the models classify nativeness of sentences. The results of the first test suite, the 8 test sets, show that the deep learning models can correctly classify nativeness in a reasonable way, and we find with the results that deep learning can obtain knowledge of not just syntactic well-formedness but plausibility of sentences. Among the models, BERT and XLNet are more correct at nativeness judgments. BERT has more strength in capturing plausibility and XLNet is better at judging well-formedness.

Secondly, the models are investigated in terms of plausibility and well-formedness, and compared with nativeness judgments of English native speakers in a prior paper. The results reveal that XLNet is the best well-formedness classifier, but the native judgments overwhelm it. On plausibility items, however, we find that BERT almost reaches the level of native judgments.

Lastly, this paper evaluates the models with test items from DeKeyser (2000). The results indicate that our models are vulnerable to cover various kinds of syntactic violations. Learnability problems regarding obtaining syntactic information are not shared between deep learning and L2 learners, which means the learning strategies of deep learning and learners are quite different. DeKeyser (2000) explains the learnability problems of learners in terms of salience. The author observes that the more salient a grammatical factor, the easily and faster the factor is learned. For example, gender error is 'perceptually salient' because '[p]ronoun gender errors are so irritating to native speakers that they will almost always correct them when their nonnative interlocutors make such mistakes, […]' (DeKeyser, 2000). This case does not occur to deep learning.

In this paper, some similarities and differences between deep learning and humans are discovered. However, still the argumentation on the cognitive underpinning that integrates learning process of L2 learners and deep learning is not clearly developed. It follows that visualization of internal vector representation is required to endorse the assumption for the future research.

References

Bley-Vroman, R. (1988). The fundamental character of foreign language learning. Grammar and second language teaching: A book of readings, 19-30.

Chaves, R. P. (2020). What Don't RNN Language Models Learn About Filler-Gap Dependencies?. Proceedings of the Society for Computation in Linguistics, 3(1), 20-30.

Da Costa, J. K., & Chaves, R. P. (2020). Assessing the ability of Transformer-based Neural Models to represent structurally unbounded dependencies. Proceedings of the Society for Computation in Linguistics, 3(1), 189-198.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. Studies in second language acquisition, 22(4), 499-533.

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

Firth, J. R. (1961). Papers in Linguistics 1934-1951: Repr. Oxford University Press.

Goldberg, Y. (2019). Assessing BERT's Syntactic Abilities. arXiv preprint arXiv:1901.05287.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

Linzen, T., & Baroni, M. (2020). Syntactic Structure from Deep Learning. arXiv preprint arXiv:2004.10827.

Linzen, T., & Leonard, B. (2018). Distinct patterns of syntactic agreement errors in recurrent networks and humans. arXiv preprint arXiv:1807.06882.

Mikolov, T., Karafiát, M., Burget, L., Č ernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In Eleventh annual conference of the international speech communication association.

Park, K., You, S., & Song, S. (2019). Using the Deep Learning Techniques for Understanding the nativelikeness of Korean EFL Learners. Language Facts and Perspectives, 48, 195-227

Park, K., You, S., & Song, S. (2020). Not Yet as Native as Native Speakers: Comparing Deep Learning Predictions and Human Judgments. English Language and Linguistics, 26, 199-228.

Pawley, A., & Syder, F. H. (2014). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Language and communication (pp. 203-239). Routledge.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In Thirteenth annual conference of the international speech communication association.

Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural network acceptability judgments. Transactions of the Association for Computational Linguistics, 7, 625-641.

Wilcox, E., Levy, R., & Futrell, R. (2019). What Syntactic Structures block Dependencies in RNN Language Models?. arXiv preprint arXiv:1905.10431.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems (pp. 5754-5764).