Korean Journal of English Language and Linguistics, Vol 21, June 2021, pp. 487-509
DOI: 10.15738/kjell.21..202106.487

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

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

http://journal.kasell.or.kr

A Deep Learning-based Understanding of Nativelikeness: A Linguistic Perspective*

Kwonsik Park (Korea University) **Sanghoun Song** (Korea University)

Received: April 25, 2021
Revised: June 5, 2021
Accepted: June 25, 2021

Kwonsik Park (1st author)
Graduate Student, Dept. of Linguistics, Korea Univ.
Tel: +82-2-3290-1648
oneiric66@korea.ac.kr

Sanghoun song (corresponding author)
Professor, Dept. of Linguistics, Korea Univ.
Tel: +82-2-3290-2177
sanghoun@korea.ac.kr

*This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020S1A5A2A03042760).

ABSTRACT

Park, Kwonsik and Sanghoun Song. 2021. A deep learning-based understanding of nativelikeness: A linguistic perspective. *Korean Journal of English Language and Linguistics* 21, 487-509.

Constructing deep learning models that identify nativelikeness in English sentences, this paper addresses two relevant research questions: is nativelikeness measurable, and is it determined by syntactic well-formedness and lexical associations? To address the first, our models are evaluated by judging every item in Test Suite I, which comprises learner and native sentences from four sources. The results show that the models predict nativelikeness reasonably well. Next, syntactic well-formedness is examined via Test Suite II, comprising correct–incorrect minimal pairs with two conditions. The results indicate that our models do not satisfactorily detect it. The learners' results reveal their limited knowledge, suggesting that the models learn the inadequateness of lexical associations as a feature of non-nativelikeness because the learner training data comprises Korean English learner corpora. However, our models' results also show poor performance. We conclude that deep learning is capable of measuring nativelikeness, and well-formedness and lexical associations are no more than necessary conditions for nativelikeness. This implies the need to consider other factors when defining and assessing nativelikeness.

KEYWORDS

deep learning, nativelikeness, well-formedness, lexical association, learner corpora

**Some earlier versions of the present study were partially presented at the 34th Pacific Asia Conference on Language, Information and Computation (online, Oct-24-2020) and the 2020 International Conference on English Linguistics (online, Oct-17-2020).

1. Introduction

While talking with non-native speakers of our native language, we can intuitively discern whether their utterances are nativelike or non-nativelike. This is because there are various degrees of L2 learners' "imperfection" resulting from apparent differences between L1 and L2 acquisition (Sorace 2003: 130).

However, as "varying definitions of 'nativelikeness' […] have had an influence on the different rates of nativelikeness reported" (Abrahamsson and Hyltenstam 2009: 291), it is hard to define nativelikeness exhaustively. In addition, more than a few factors are involved in judging nativelikeness (Abrahamsson and Hyltenstam 2009, Pawley and Syder 1983). Furthermore, Birdsong (2005: 321) posits "a basic problem of applying the non-native-likeness standard: deciding what is and is not appropriate evidence for deficits in the learning mechanism." No clear criteria of nativelikeness, consequently, results in no standardized way of assessing nativelikeness (Bylund et al. 2012, Schimitt 1998).

According to Abrahamsson and Hyltenstam (2009: 259), "[w]hether someone is perceived as a native speaker by (actual) native speakers […], is a central aspect of nativelikeness." The attainment of nativelikeness requires comprehensive linguistic knowledge including phonological, morphological, syntactic, and semantic proficiency (Long 1990). Other factors such as the knowledge of interactional sociolinguistics and discourse-level contextual relevance are also needed (Magnusson and Stroud 2012, Pawley and Syder 1983).

Several studies addressing nativelikeness propose their own standards to define and measure it. For example, Clahsen and Felser (2006b) focus on investigating learners' processing of grammar information, such as long-distance dependency. In a broader perspective, Guerssel et al. (1985) point out that native speakers have competence of uttering "syntactically relevant semantically coherent verb classes", which means, for a sentence to be nativelike, not only does it have to be syntactically adequate, but it also accords with general world knowledge. Similarly, Pawley and Syder (1983: 191), emphasizing "institutionalized" and "lexicalized" patterns, discuss how a native speaker "selects a sentence that is natural and idiomatic from among the range of grammatically correct paraphrases." However, we cannot definitely determine how nativelike a sentence is with those standards. Despite the vague concept, nativelikeness has been importantly studied as we can use natives' judgments as a means of detecting errors of learners' judgments by comparative analyses, and furthermore, as Birdsong (2005: 320) states that "the standard by which the upper end of L2A [second language acquisition] attainment is typically measured is nativelikeness."

To scrutinize the concept of nativelikeness, this study chooses the strategy of constructing nativelikeness judgment models with state-of-the-art deep learning technology. We first address the question of, (i) whether deep learning can assess nativelikeness, i.e., whether nativelikeness is measurable. Subsequently, based on deep learning's judgment behaviors in the first experiment, we pose another question, (ii) whether syntactic well-formedness and lexical associations—linguistics factors of the sentence unit—are decisive factors that influence nativelikeness, and how influential they are.

Deep learning has become a useful tool to conduct scientific research on human language (Linzen 2019). Previous studies have evaluated deep learning models largely by examining whether they have learned any syntactic information, such as filler-gap dependencies (e.g., Wilcox et al. 2019), subject-verb agreement (e.g., Marvin and Linzen 2018), and reflexive anaphora (e.g., Linzen and Leonard 2018). Quite a few previous studies have proven that deep learning can obtain syntactic knowledge from language data thus far (Warstadt et al. 2020, *inter alia.*). This remarkable performance may suggest that we are gradually nearing the goal of replicating humans' cognitive-linguistic abilities with the advanced technologies.

The erstwhile studies commonly argue that it is feasible for deep learning machines to model human language

processing with native data. We assume, by extension, that deep learning with learner data can also progress in language representations in that learner data represents what native data cannot (i.e., non-nativelikeness). Proving whether utilizing learner data is beneficial to building language models could enrich our theories on natural language processing (NLP) with deep learning. Accordingly, the present work attempts to examine the feasibility of using native plus learner data to represent humans' judgment standards for nativelikeness.

More specifically, this work implements deep learning models that can detect the nativelikeness of English sentences to see whether we can explore a more superordinate concept than syntax with deep learning. Accordingly, this paper chooses four kinds of deep learning networks trained with native and learner corpora, and then each model is evaluated by judging the nativelikeness of every sentence in three test suites to examine the judgment aspects from various angles. Successful implementation of a nativelikeness detector would demonstrate that nativelikeness is measurable. The measurability would then shed light on which factors exert strong influences on making a sentence sound nativelike as we can investigate nativelikeness by examining why a model judges a sentence as nativelike or not.

This paper mainly focuses on substantiating the measurability of nativelikeness as it should precede deeper investigation on the concept from more various linguistic perspectives. Proving objectivity of the methodology proposed in the current work is expected to be a cornerstone for quantitative research on nativelikeness with deep learning.

2. System

2.1 Data

The entire data, comprising native and learner texts, amounts to 651,666 sentences. Of this, 90% consists of training data. And 10% consists of validation data (randomly excerpted), which is used to evaluate the judgment accuracy of the deep learning models. The learner data is excerpted from the Yonsei English Learner Corpus (YELC; Rhee and Jung 2012) and the Gachon Learner Corpus (GLC; Carlstrom and Price 2013), made by undergraduate students at Yonsei University and Gachon University, South Korea, respectively. The native data is extracted randomly from the Corpus of Contemporary American English (COCA; Davies 2008) to form the 50:50 proportion (native: learner), so that the native and learner sentences are 325,833, 325,833, respectively.

Every sentence in the training data is labeled "0" or "1," the former indicating learner data and the latter, native data. During the training, the models generalized the features of sentences labeled "0" and those labeled "1." Subsequently, they were evaluated by predicting whether a sentence is close to "1" or "0." If the output score of an input sentence was "0" or closer to "0," we regarded it to be judged as a non-native sentence by the model, and if it was "1" or closer to "1," as native.

The current analysis employs the dichotomous classification between native and non-native judgments, although we posit gradience in grammar (Sorace and Keller 2005; and others). This is largely because such simple either-or judgments are also widely exploited and preferred in experimental syntax studies, providing statistical evidence for syntactic phenomena (Myers 2009; Sprouse et al. 2013; Song and Oh 2017).

2.2 Model

The present work constructs four language models using different deep neural networks, which are proven to

learn syntactic knowledge. Using the following four deep learning networks, this research investigates whether the models can detect nativelikeness in English sentences. The following are brief explanations of the models (more detailed information is available in Krohn et al. (2019); Bengio et al. (2017); Goodfellow et al. (2016); etc.).

Recurrent Neural Network (RNN; Rumelhart et al. 1986): A type of artificial neural network in which information from a previous step is updated in the current step, so that the memorized history of words is used as a context in resolving a task, such as next word prediction.

Long-Short Term Memory network (LSTM; Hochreiter and Schmidhuber 1997): It resolves to some extent the problem of gradient vanishing by updating information from previous steps selectively; what is considered unimportant is discarded. This leads to improved memorization efficiency and thereby the network performs better than RNN.

Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. 2018): 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: 1). BERT is powerful enough to outperform many previous NLP models, but is also limited as it "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: 2).

XLNet (Yang et al. 2019): Yang et al. attempted to elaborate BERT to build XLNet, which is capable of conducting both autoregressive and autoencoding methods by "[maximizing] the expected log likelihood of a sequence w.r.t. all possible permutations of the factorization order."

2.3 Machine

We trained each model for 10 epochs using a GPU (GeForce RTX 2080 Ti Dual D6 11GB) and the best model with the highest accuracy was saved and used for nativelikeness judgments. BERT and XLNet models (BERT-Base and XLNet-Base versions, respectively) were trained for 4 epochs (the BERT authors recommend 2 to 4 epochs when finetuning a model) with the same GPU. We used a batch size of 32 for all four models. PyTorch library was used to implement the networks (Paszke et al. 2017).

We also deal with deep learning's *black box* problem of neural networks (Alain and Bengio 2016), which means the learning processes of deep neural networks are so intricate that we cannot interpret the language representations of the human brain, by utilizing a visualization technique or *heatmap*. In the present study, heatmaps are pictured based on how much *attention* a model gives to each word in a sentence. Attention is an embedding technique that allocates contextual information to words by computing weighted averages for them (Vaswani et al. 2017). It facilitates seeing which word a model pays attention to when predicting the nativelikeness of a sentence.

3. Experiment I: Is Nativelikeness Measurable?

3.1 Goals

This experiment aims at confirming the feasibility of implementing deep learning models that adequately predict the nativelikeness of English sentences. Investigating factors that influence nativelikeness must be

preceded by constructing nativelikeness detectors to facilitate the analyses of the factors. Furthermore, the present experiment is necessary as the prediction results can help uncover the factors of nativelikeness; we can roughly estimate which factors might have influenced the results.

This experiment uses native and learner data for the networks to represent nativelikeness and non-nativelikeness, respectively. Distinguishing nativelikeness from non-nativelikeness judgments is carried out here based on whether a sentence is written by a native speaker or a learner. We attempt to model the way L2 learners learn English (not L2 itself). Because BERT and XLNet are pre-trained networks with native English data and already have embedded information about English, BERT and XLNet represent native English teachers. As RNN and LSTM have no pre-embedded information, RNN and LSTM represent non-native English teachers.

3.2 Methods

Warstadt et al. (2019) construct their classification models trained with the Corpus of Linguistic Acceptability (CoLA), comprising English sentences labeled with a binary grammaticality value (data sources are published linguistics literature). Our model training differs from previous work as we train models with learner data, which has grammatical as well as ungrammatical sentences. We employ learner data for two reasons. First, it is more reasonable to use pure, raw learner data because the present work intends to let deep learning learn non-nativelikeness from the learner data as it is. Second, it is quite tricky to reserve large amounts of ungrammatical data; for example, CoLA is proper for models to learn the unacceptability of sentences but contains only 10,657 sentences (Warstadt et al. (2019) used 8,551 examples in CoLA to train their model), which seems too small to be used as training data. One could use an artificially created error corpus (e.g., Rozovskaya and Roth 2010). However, we do not resort to this strategy because such a corpus cannot bear all the characteristics of naturally occurring learner data; it does not reflect actual error patterns committed by the learners.

3.2.1 Test data: Test Suite I

Test data is similar to validation data as both are used to evaluate a model's performance; but it also differs as test data comprises external data that is not excerpted from the training data, whereas validation data has the same characteristics as training data. As shown in Table 1, four test sets (collectively called Test Suite I here) are chosen to probe whether the models correctly judge the nativelikeness of English sentences. Each test set has a representative characteristic. They are as follows: (i) English Gigaword (Graff et al. 2003) is a highly elaborated version of native English data comprising 3,000 sentences randomly excerpted from original data constructed from news data (native data); (ii) the Speckled Band is novel data characterized by the inclusion of many daily and easy expressions (native data); (iii) the Tanaka Corpus (Tanaka 2001) comprises 3,000 sentences randomly extracted from original data of English learners' sentences. As the researchers edited it during corpus construction, it lacks syntactic violations, which means it has the characteristics of both learner and native data (learner/native data). The last one, (iv) the written version of the Incheon National University Multilingual Learners Corpus (hereafter, INU-MULC; Yoon et al. 2020) was written by 34 Korean, 1 Uzbek, and 1 Persian learners of English (learner data). Characterized by its contents of daily, casual conversations, it is a pure, raw learner corpus with large quantities of syntactic errors. The two native data sets are chosen to compare the influences of formal versus informal data on nativelikeness, and the learner/native and learner data sets are used to compare nativelikeness of syntactically well-formed versus ill-formed data.

Table 1. The 4 Test Sets (Test Suite I)

Type	Subtype	Test Data	Number of Sentences	Type-Token Ratio
Native	News	English Gigaword	3,000	0.1968
Native	Novel	Speckled Band	572	0.2113
Learner/Native	Edited	Tanaka Corpus	3,000	0.1621
Learners	Not edited	INU-MULC	613	0.1775

3.3 Results

The validation accuracies of the four models are 94.27% (RNN), 95.35% (LSTM), 97.82% (BERT), and 97.33% (XLNet). The models are also evaluated using Test Suite I by judging nativelikeness for every sentence in each test set. Numbers in Figure 1 are the proportions of sentences that are predicted as native in each test set. A higher score means better performance for native data whereas a lower score means better performance for learner data. As shown in Figure 1, BERT and XLNet perform better than RNN and LSTM. XLNet outperforms the rest at predicting native test sets and BERT at non-native test sets.

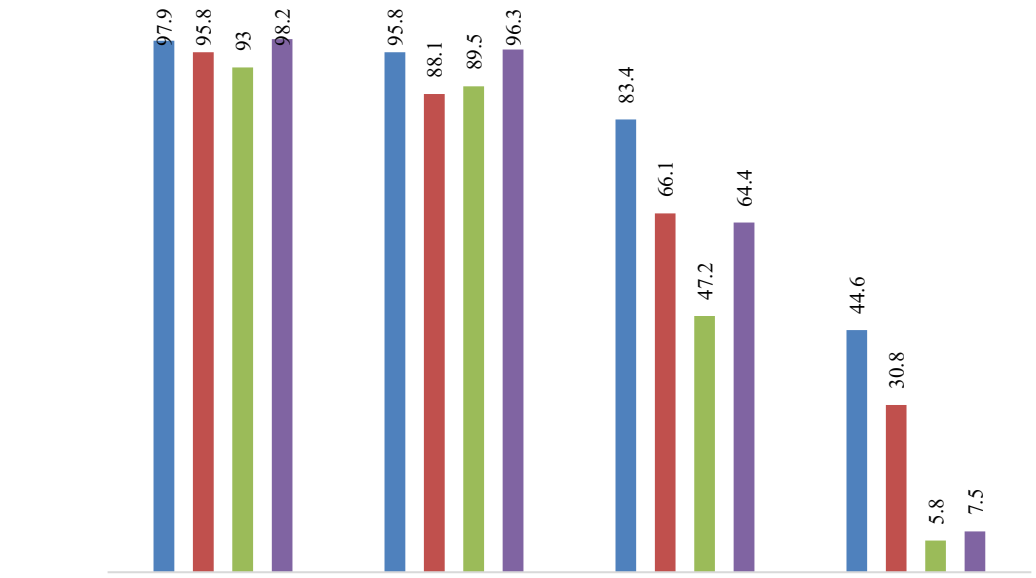


Figure 1. Nativelikeness Judgment Results for Test Suite I

For the native test sets, the four models are able to discriminate between native and non-native sentences. The following sentences are excerpted from the native test sets.

(1) a. He favored *provincial autonomy* and award of more shares for provinces. (Gigaword)

b. That was the only *deterrent*, he said. (Gigaword)

(2) a. This is very interesting. (Speckled Band)

b. That is important. (Speckled Band)

Here, (1a) and (1b) are judged as native sentences by all the models. *Provincial autonomy* in (1a) is a technical term rather than an informal one; *deterrent* does not commonly appear in English data and is not included in the top 5,000 words in COCA (Davies 2013). Meanwhile, (2a) and (2b) are predicted as non-native by all four models. They contain commonly used words that learners also use frequently. Learners are fully capable of producing such sentences.

We attempted to identify why the models judge a sentence as native or non-native using heatmaps. Figure 2 shows the attention scores that BERT, the highest performing model, gives to each word in (1a); the higher attention score a word has, the stronger impact it has on a nativelikeness judgment. Note that a brighter color represents more attention paid to the word when models predict a sentence as native or non-native.

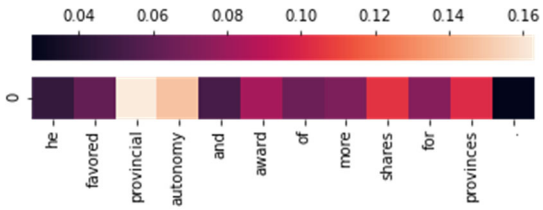


Figure 2. Heatmap Analysis I (sentence: *He favored provincial autonomy and award of more shares for provinces.*)

As shown in Figure 2, the term *provincial autonomy* is given the highest attention scores, which means it influences BERT's nativelikeness prediction the most. This reveals that infrequent and/or difficult technical words are likely to affect sentences' nativelikeness.

Regarding the non-native test sets, the Tanaka Corpus and INU-MULC, the proportions of sentences predicted as native are lower than those of native test sets. However, the results of all the models except for BERT on the Tanaka Corpus are more than 50%, which means the three models judge the overall characteristics of the Tanaka Corpus to be nativelike rather than non-nativelike. In contrast, INU-MULC is judged appropriately as non-native data by all models. A critical difference between the Tanaka Corpus and INU-MULC is whether they have syntactically ill-formed sentences. Meanwhile, they are similar regarding lexical choices. The following are some sentences predicted as non-native by all models.

(3) a. My uncle gave me a book. (Tanaka)

b. My father is proud of my being handsome. (Tanaka)

(4) a. It's actually school homework. (INU-MULC)

b. *I worried and tired because a lot of people. (INU-MULC)

Here, (3a) and (3b) are syntactically well-formed but predicted as non-native partially because their lexical choices are learner-like rather than nativelike; (3a) follows a pattern that mostly appears in learner textbooks, and (3b) sounds awkward due to its unidiomatic turn of phrase, similar to "I wish to be wedded to you" (Pawley and Syder 1983: 191); it could be better written with a simpler and easier sentence structure but is used in an overly

grammatical one. Like (3b), (4a) does not read naturally.

Although native speakers might produce syntactically erroneous utterances (Ivanova et al. 2012), determining whether a sentence has any syntactic violations seems like a reasonable way to weed out learner sentences. It is a natural assumption that syntactic violations occur much more frequently in learner data. Here, (4b) contains the syntactic violations of omitting the *be*-verb and using *because* instead of *because of*. Sentences like (4b) are definitely considered learner sentences due to syntactic ill-formedness. Figure 3 represents the attention scores of each word in (4b).

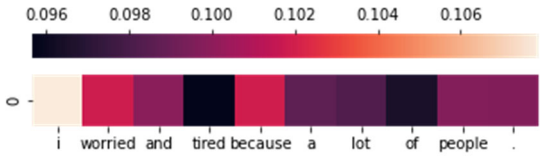


Figure 3. Heatmap Analysis II (sentence: **I worried and tired because a lot of people.*)

As shown in Figure 3, BERT pays relatively higher attention to the syntactically ill-formed tokens, *worried* and *because*. Meanwhile, the first element *I* gets the highest attention, which can be explained by BERT's tendency to give higher attention to initial tokens. This is quite reasonable as humans also pay more attention to the first word in a sentence because it normally entails much information. From the perspective of information structure, the subject argument is generally considered a "default" topic, which "contextualize[s] other elements in the clause" (Goldberg 2006: 130), revealing the "current concern" of a sentence (Lambrecht 1986: 100).

3.4 Discussion

It can be argued that deep learning would be useless if the mechanisms of it are not discernible. However, although countless AI techniques with deep learning networks have been developed without understanding their inner mechanisms clearly, this should not be a reason to suspend studying linguistics with deep learning. Instead, we should endeavor to delve further into and understand the inner mechanisms of artificial neural networks to take full advantage of deep learning. This paper, attempting to resolve the black box problem, exploits heatmap analyses to figure out where a model pays much attention on a token in a sentence when judging nativelikeness.

Given the accuracy of nativelikeness judgments by the four models, the nativelikeness of sentences is measurable to a certain degree. This resolves the first research question of whether deep learning can learn nativelikeness. Note that this article chooses the four models not to compare them but because each has been proven to be able to learn a language. In addition, the four models are used to confirm if deep learning technologies are developing in the right direction; the fact that the ability of BERT and XLNet, relatively up-to-date networks, to predict nativelikeness improves drastically indicates that the development of deep learning is on the right track (see Figure 1). Thus, studying nativelikeness with deep learning seems valid.

For the native test sets, the four models predict their nativelikeness accurately in general. All models give English Gigaword (average 96.23%) higher scores than the Speckled Band (average 92.43%). This is probably because English Gigaword (news data) consists of more technical and/or not commonly used words, which do not appear frequently in learner training data, whereas the Speckled Band (novel data) includes more commonly used daily and easy words. However, we are not asserting that the frequency of the use of daily and easy words is the only critical factor that determines nativelikeness as the average score of English Gigaword is not considerably different from that of the Speckled Band, but arguably it contributes to predicting nativelikeness.

Regarding the non-native test sets, the proportions of what are predicted as native sentences are lower than those of native test sets. However, the results reveal that the Tanaka Corpus is considered more nativelike than non-nativelike. Nonetheless, the scores imply that the models judge the Tanaka Corpus quite reasonably. One reason could be that the Tanaka Corpus is a manually edited version of its original corpus collection and has no sentences with syntactic violations. INU-MULC, in contrast, is predicted appropriately as non-native data; it has the most learner-like data in Test Suite I, as it is pure, raw learner data which has not been edited at all.

In sum, the models detect nativelikeness as intended. The next step is to scrutinize where the nativelikeness is in a sentence. Given the results for predicting the two non-native test sets, we can estimate that one factor determining nativelikeness is syntactic well-formedness; one reason more sentences in the Tanaka Corpus are predicted as native than in INU-MULC is that the Tanaka Corpus does not have syntactically ill-formed sentences. This leads us to conclude that deep learning models can correctly detect syntactic information in English sentences.

4. Experiment II: Is Syntactic Well-formedness a Decisive Factor?

4.1 Goals

In this section, we test whether deep learning models can distinguish between syntactically ill-formed and well-formed sentences. Experiment I validates that deep learning can measure the nativelikeness of English sentences. Presuming that nativelikeness exists somewhere in a sentence, we next attempt to pinpoint its location. Given Experiment I's results, we hypothesize that syntactic well-formedness is one factor that critically influences nativelikeness.

4.2 Methods

4.2.1 Test data: Test Suite II

To examine if the models have acquired syntactic knowledge, we chose syntactic well-formedness test items from DeKeyser (2000), a revised version of those from Johnson and Newport (1989). The original test items comprised 200 sentences including four pretest items. Excluding the pretest items, there are 98 correct–incorrect minimal pairs with two conditions, grammatical and ungrammatical. Categorized into 11 types (27 subtypes), the test set is a highly refined set of items to measure test takers' knowledge of diverse 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 lexical associations (which are discussed in the next section). This research conforms to the standards of selecting items and classifies them into (sub)categories in DeKeyser (2000). We judge that the test items are reliable to test syntactic knowledge as (i) they include a lot of syntactic categories, and (ii) the paper is one of the recognized articles in testing learner's grammaticality. The 11 types for testing the knowledge of well-formedness are exemplified in (5). A brief list of 27 subtypes is given in Appendix A.

(5) a. Last night the old lady [died/*die] in her sleep. (past tense)
b. Three [boys/*boy] played on the swings in the park. (plural)
c. John's dog always [waits/*wait] for him at the corner. (third-person singular)

Using the results for nativelikeness judgment for as many as 27 subtypes enables us to investigate the models' syntactic knowledge in a comprehensive perspective. We predict nativelikeness for every sentence in Test Suite II with the four models.

4.3 Results

Table 2 shows the results for Test Suite II, counted in two units: (i) discrete sentences, and (ii) minimal pairs. The number of discrete sentences in Table 2 indicates how many sentences are correctly predicted, and minimal pairs represent how many sentence pairs are predicted correctly. As seen, XLNet outperforms the rest.

Table 2. Results for Test Suite II

	RNN	LSTM	BERT	XLNet
Discrete Sentences (196 items in total)	99 (50.51%)	91 (46.43%)	104 (53.06%)	**114 (58.16%)**
Minimal Pairs (98 pairs in total)	4 pairs	4 pairs	14 pairs	**21 pairs**

In addition, we use heatmap pairs that are predicted correctly by BERT and XLNet to confirm whether the error tokens we designed are given the highest attention. Figure 4 shows some comprehensible examples whereas Figure 5 includes sentences for which we cannot understand the reasons for the respective prediction. Here, 11 out of 14 pairs that BERT predicts correctly are comprehensible, i.e., error tokens get relatively higher attention. However, only 2 out of 21 pairs that XLNet answers correctly are interpretable, indicating it is difficult to explain how XLNet gets the right answers to the prediction task. Note that we cannot analyze predictions of our RNN and LSTM models with heatmaps because (i) they do not use the attention mechanism, and (ii) in RNN and LSTM networks' algorithms, weights are updated consecutively and cumulatively from the first token to the last, with each weight changing depending on the prior one, which means that each word has integrated information for itself and its precedents, rather than individual information for that word.

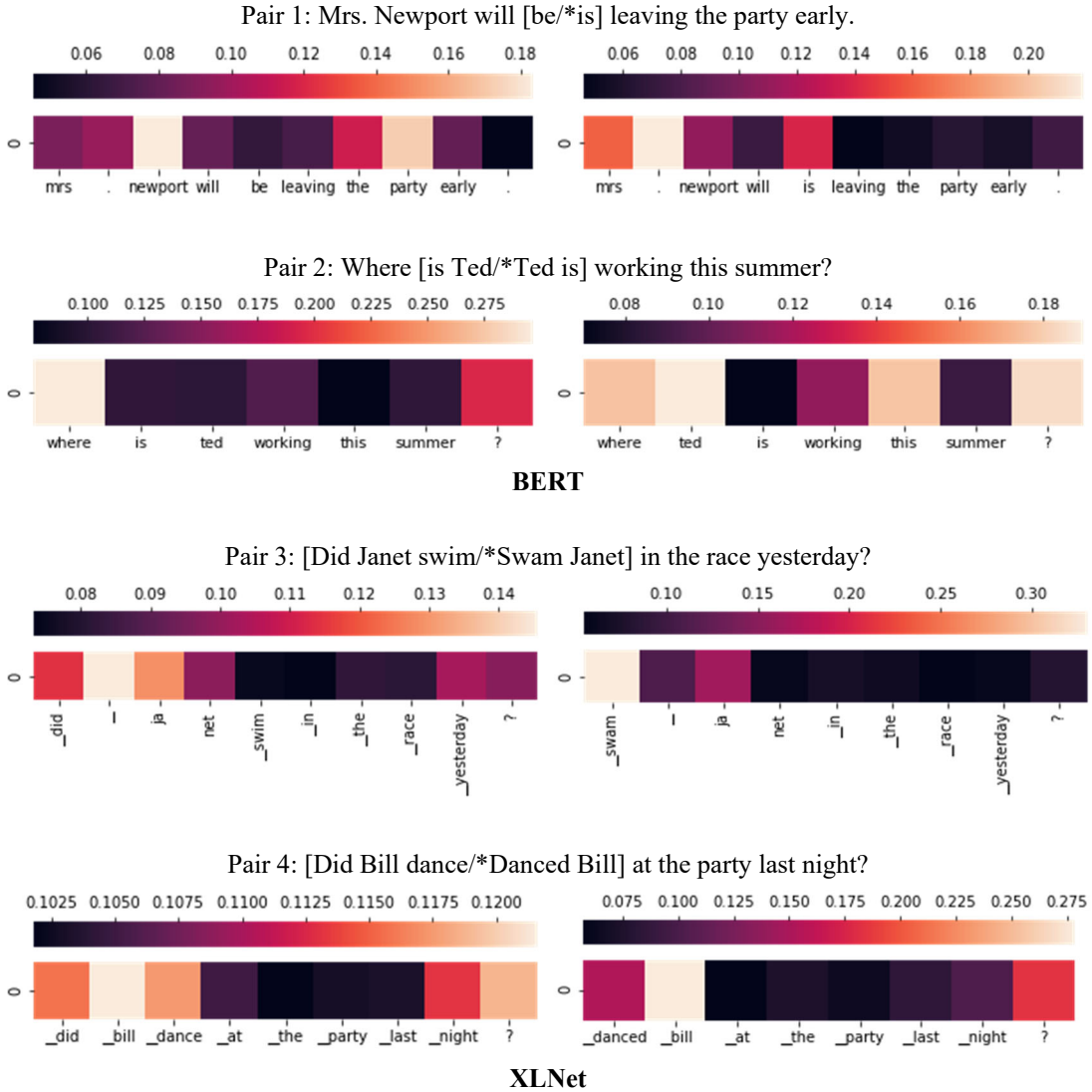
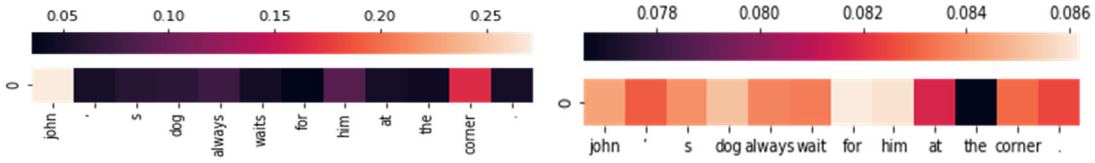


Figure 4. Heatmap Analysis III (pairs that are predicted correctly by BERT and XLNet)

In Pair 1, there is an ill-formed expression such as *is*. Note that *will is* on the right side is brighter than *be* on the left. *Ted* in Pair 2 is the brightest on the right side but dark on the left. In addition, *swam* in Pair 3 and *Bill* in Pair 4 are the brightest. Figure 4's heatmaps show that words that deserve attention do receive more. This confirms that well-formedness acts as a significant factor for some test items.

Pair 5: John's dog always [waits/*wait] for him at the corner.



Pair 6: Last night the old lady [died/*die] in her sleep.

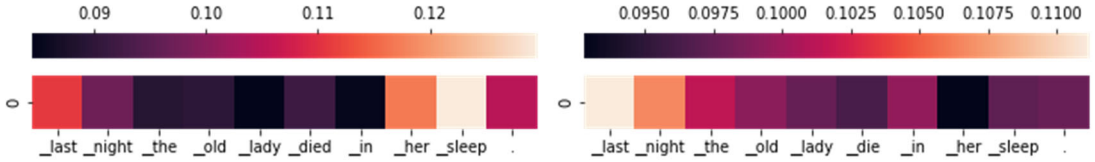


Figure 5. Heatmap analysis IV (pairs that are predicted correctly by BERT and XLNet)

Like Figure 5, however, there are also items we cannot comprehend how they judge the pairs correctly.

4.4 Discussion

Some predictions show BERT has somewhat learned syntactic information and attention is paid reasonably in some heatmaps; the fact that some heatmaps are interpretable on the basis of syntactic well-formedness means we cannot ignore the role of syntactic constraints in certifying nativelikeness. Simultaneously, however, the fact that not all heatmaps are explicable indicates that we should consider other factors involved in defining nativelikeness. In addition, considering the high nativelikeness prediction accuracies of all four models in Experiment I, the correct rates in Experiment II are substantially low. This means not only they have failed to fully learn syntactic information, but also syntactic well-formedness is not the one critical factor which affects nativelikeness judgment. This is somewhat reasonable as learner sentences also include syntactically well-formed sentences.

To summarize, Experiment II's overall results proffer that syntactic well-formedness is not a sufficient but just one necessary condition for nativelikeness. This rejects the first attribute of the second research question. Note that we do not claim that syntactic knowledge is not an important factor influencing nativelikeness; we argue that it is not a decisive factor that determines a sentence's nativelikeness.

We next investigate another linguistic factor: lexical associations. According to Schmitt (1998), word associations can also be regarded as a significant factor in determining nativelikeness. The next section aims to confirm whether lexical associations contribute critically to predicting the nativelikeness of sentences.

5. Experiment III: Is Lexical Association a Critical Factor?

5.1 Goals

Previous research argues that the knowledge of lexical associations can accompany a kind of learnability problem, so the knowledge is indispensable to being a nativelike speaker (Nesselhauf 2003; Wray 2000).

This experiment is carried out with our deep learning models and Korean English learners. The four models

and participants judge the acceptability of the same sets of items (Test Suite III). We conduct human evaluation to investigate whether Korean English learners lack the knowledge of lexical associations. As the training data was made by Korean learners of English, we presume that if participants are inept at identifying inapposite lexical associations, such as "*powerful/#strong computers*" (Leacock et al. 2010: 27), the models would have been trained with data comprising a host of inappropriate lexical associations. If it is the case, it is likely that our models have learned (non-)nativelike lexical associations as a critical feature to distinguish between native and non-native sentences.

We follow the definition of a lexical association as proposed by the articles from which we extracted the test items: (i) It is a collective behavioral pattern of certain words, also referred to as "conventional word combinations, or collocations" (Leacock et al. 2010: 27). (ii) It is characterized by the property of "non-substitutability," which means "the inability to substitute a near synonym for part of a collocation" (Leacock et al. 2010: 65). That is, "the habitual co-occurrence of two or more words whose meaning can be inferred from the parts, but will become less acceptable when one of the elements is replaced by a similar word" (Shei and Pain 2000: 1). (iii) Furthermore, collocations can also be characterized by their "arbitrary restriction on substitutability": if a restriction on certain word combinations is to some extent arbitrary, the combinations are called collocations (Nesselhauf 2003: 255).

5.2 Methods

5.2.1 Experimental setting

For the current experiment, 85 native Korean learners of English from diverse affiliations were recruited. Their ages ranged between 19–55 years. Their mean age was 30.14 (*SD* = 9.04). The percentage of females and males was 64.39% and 35.71%, respectively (Those who did not respond with their personal information were not counted). This experiment was carried out using the online survey management tool *Survey Monkey*. For Test Suite III, subjects responded based on goodness of the stimuli because it is unreasonable to ask non-natives to judge the degree of nativelikeness. The experiment was implemented using a 5-point Likert scale (1 = *very bad* and 5 = *very good*), and the responses were transformed into *z*-scores for analyses to avoid potential bias. The stimulus sentences in Test Suite III were presented on a screen one by one in random order.

5.2.2 Test data: Test Suite III

The stimulus items comprise two parts: (i) the Michigan Test (Ko et al. 2010), and (ii) Test Suite III. Replicating Ko et al. (2010), we used the Michigan Test (comprising 30 multiple-choice items) to assess the participants' English proficiency. The subjects were divided into four subgroups based on the Michigan Test scores: 20 beginners (0–12), 21 low-intermediate learners (13–17), 24 high-intermediate learners (18–22), and 20 advanced learners (23–30). In all, 39 subjects carried out the Michigan Test before taking Test Suite III, and 46 before taking the main test suite. Each subject answered four pretest stimuli at the beginning of the experiment to familiarize themselves with the main experiment.

Test Suite III comprises well-formedness and lexical association judgment test sets (well-formedness items are included to compare results of the well-formedness test with those of the lexical association test). Well-formedness test items are excerpted from a textbook exploring automated grammatical error detection technologies (Leacock et al. 2010). The textbook includes various types of errors committed by English learners.

Various categories of syntactic errors are intended to examine participants' syntactic knowledge comprehensively. To make the lexical association test set, we extracted error expressions that appear in the textbook and two articles cited therein (Nesselhauf 2003; Shei and Pain 2000). Whereas well-formedness test items are all excerpted from the textbook, we additionally chose the two articles addressing collocations in second language learning to build the lexical association test items, as the textbook includes very few examples for lexical associations. The two types of test items are illustrated below. Well-formedness errors are categorized into 11 types and lexical test items are uncategorized. Words that are not included in the top 5,000 in COCA (Davies 2013) were not selected to avoid cases where subjects answer wrong simply because they do not know a word in a sentence. Many error patterns are not presented as a full sentence, such as, say, *eat the medicine* instead of *Tom ate the medicine*. Thus, we manipulated such expressions into full sentences. Some sentences in Test Suite III are exemplified in (6). A brief list of items from Test Suite III is given in Appendix B.

(6) a. My best friend [knows/*know] this guy. (Well-formedness)
b. I want to make [myself/*me] fit. (Well-formedness)
c. We [do not have/*have no] any time. (Well-formedness)
d. The company bought a [powerful/#strong] computer. (Lexical associations)
e. Tom [took/#ate] the pill. (Lexical associations)
f. Mary wants to [shoot/#take] a film. (Lexical associations)

Test Suite III items are structured in the Latin square design (Gao 2005) to avoid participants becoming aware of the intention of stimuli due to similar patterns for two sentences in a minimal pair. For example, the participants could have noticed the intention of stimuli if two items of (7a) had occurred in the same experiment. Thus, we made two test settings: Test Set A including (7a-*powerful*) and (7b-*#strong*), and Test Set B including (7a-*#strong*) and (7b-*powerful*). The two settings are thereby re-lexicalized minimal pairs.

(7) a. The company bought a [powerful/#strong] computer.
b. The factory produced a [powerful/#strong] machine.

The total number of sentences in Test Sets A and B is 168. The ratio between well-formedness and lexical association test sets is 50:50. Test Set A (84 sentences) was administered to 39 and B (84 sentences) to 46 participants.

5.3 Results

Table 3 shows the results of judgments by the models and participants on the same stimulus items.

Table 3. Nativelikeness Judgment Results for Test Suite III

Agent	Level/ Network	Correct rate									
		well	¬well	tot. (well)	lex	¬lex	tot. (lex)	tot. (all)	pair (well)	pair (lex)	tot. (pair)
Human	Beginner	47.7	47.6	47.6	38.1	**54.8**	46.5	47.1	26.2	**23.8**	25.0
	Low-Intermediate	59.5	69.1	64.3	69.1	28.6	48.8	56.6	35.8	21.4	28.6
	High-Intermediate	69.1	85.7	77.4	83.4	28.6	56.0	66.7	54.8	**23.8**	39.3
	Advanced	**83.4**	**95.3**	**89.3**	**88.1**	26.2	**57.2**	**73.2**	**78.6**	21.4	**50.0**
	Total	64.9	74.4	69.6	69.7	34.6	52.1	60.9	48.8	22.6	35.7
Deep Learning	RNN	**66.7**	28.6	47.6	**66.7**	40.5	**53.6**	50.6	**26.2**	19.1	**22.6**
	LSTM	50.0	50.0	50.0	61.9	45.3	**53.6**	51.8	11.9	21.5	16.7
	BERT	21.5	**85.7**	53.6	21.5	**83.4**	52.4	53.0	11.9	**26.2**	19.1
	XLNet	42.9	71.5	**57.2**	52.4	52.4	52.4	**54.8**	21.5	16.7	19.1
	Total	45.3	58.9	52.1	50.6	55.4	53.0	52.6	17.9	20.8	19.4

***** Each number in the cells indicates the number of correct answers as a percentage. well/¬well = well-formed items/ill-formed items; lex/¬lex = lexically well associated items/lexically badly associated items; pair = the number of pairs both judged correctly; tot. = total.

As shown in Table 3, the results for Test Suite III are classified into two categories: (i) test results for syntactical well/ill-formedness, lexically good/bad association, subtotal of well-formedness/lexical association, a total of both syntactic well-formedness and lexical association, and (ii) test results for pairwise syntactic well-formedness/lexical association test items, and total pair items. We estimate that an agent can judge the well-formedness and/or lexical association well only if the agent correctly judges both good and bad items for each test set. For example, we cannot say BERT predicts well-formedness correctly, as it does not predict syntactically well-formed sentences well (21.5%) despite its relatively correct prediction rate for ill-formed items (85.7%).

Regarding human judgments, as seen, the correct rates for syntactic well-formedness items (total) increase with the level of subject groups. The correct rates of pairwise syntactic well-formedness test items also increase depending on the proficiency level. Conversely, regarding the lexical association test set, the correct rates of lexically badly associated items are considerably low among the groups (except Beginners). The correct rates of lexically well associated items are quite high compared to bad items (except Beginners), which means that the participants, even advanced learners, do not consider the adequateness of lexical associations, focusing more on syntactic information when judging a sentence. On the one hand, we cannot conclude that Beginners are better at detecting unidiomatic lexical association, because the score of lexical association items (total) does not support that: the score is the lowest (46.5%). Furthermore, the low scores of pairwise lexical association test items (average 22.6%) imply that the English learners generally have little knowledge of appropriate lexical associations; it also means that our models, trained with corpora made by Korean learners of English, have likely learned the inadequateness in collocating words as a critical feature of non-nativelikeness.

However, our models do not seem to have learned lexical associations thoroughly. Regarding the results of deep learning judgments, the correct rates of pairwise lexical associations test items are substantially low.

Table 4 shows how many minimal pairs are given the same predictions by each model. For example, two items of each minimal pair, (8a) and (8b), are given the same predictions by all models, as non-native sentences.

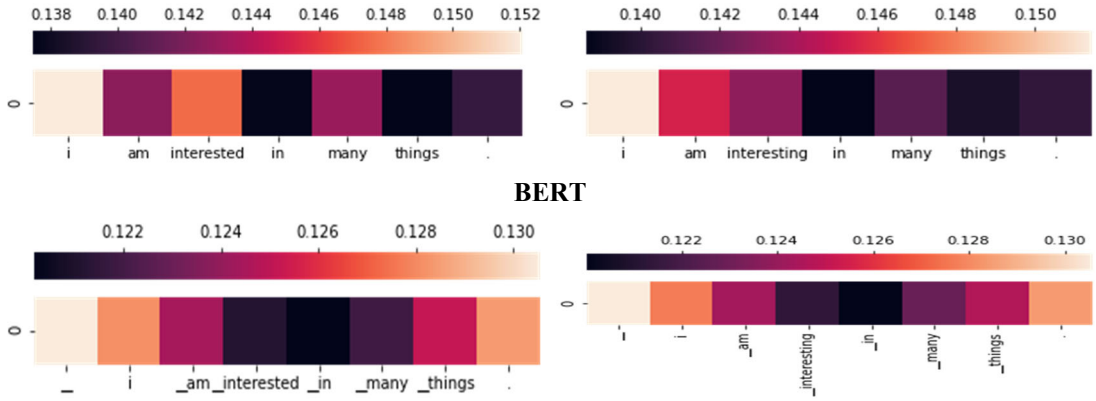
(8) a. I am [interested/*interesting] in many things.
b. The company bought a [powerful/#strong] computer.

Table 4. Number of Minimal Pairs Given the Same Predictions by Each Model

RNN	LSTM	BERT	XLNet
76.19%	75.00%	84.51%	**85.72%**

As shown in Table 4, a number of minimal pairs are given the same predictions, which suggests the models, especially BERT and XLNet, see other factors in sentences rather than each targeted correct and incorrect word we designed. Heatmaps of (8a) by BERT and XLNet support this (see Figure 6): in case of (8a), heatmaps are drawn quite similarly because from the deep learning perspective, each sentence in the minimal pair (8a) is differentiated merely by shifting one word, i.e., *interested*/**interesting* is not a game-changing exchange. In other words, the words do not play a crucial role in judging nativelikeness and it is reflected in the same responses to each sentence in the minimal pair.

Pair 1 (8a): I am interested/*interesting in many things.



Pair 2 (8b): The company bought a powerful/#strong computer.

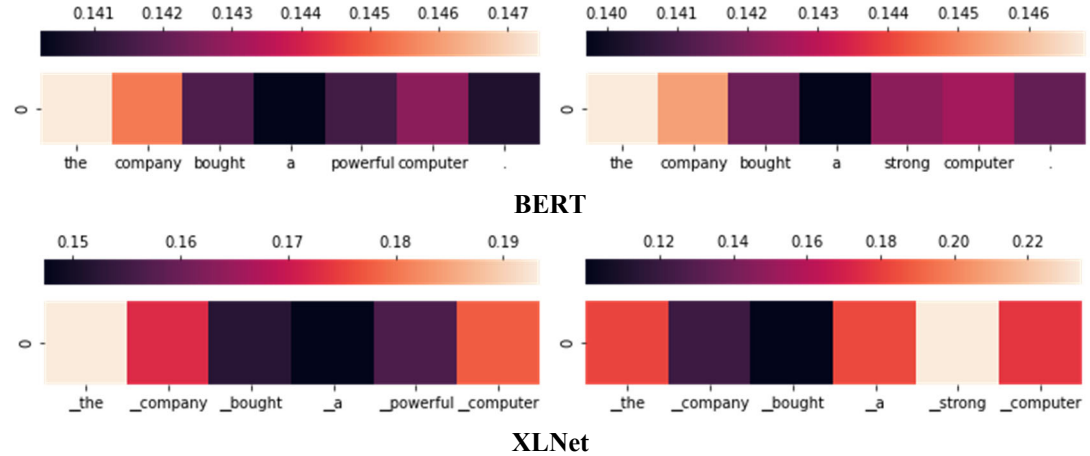


Figure 6. Heatmap analysis V (minimal pairs that are given the same predictions by BERT and XLNet)

As to Pair 2, BERT pays attention in a similar way, but XLNet attends to bad and good lexical associations differently. For the bad association, it pays the greatest attention to *strong,* the error we designed; for XLNet, a lexically bad association is a crucial factor influencing its judgment. However, for the good association, it pays more attention to *the*, *company*, and *computer,* which we cannot explain.

5.4 Discussion

As expected, the human evaluations substantiate that Korean learners lack knowledge of lexical association regardless of their English proficiency. However, given our models' poor performance on lexical association items, they do not seem to have learned the idea of the inadequateness of lexical associations as a critical factor for non-nativelikeness. This rejects the second attribute of the second research question, too, and proves that other factors of nativelikeness should be considered.

The fact that a number of minimal pairs are given the same predictions hints to us that nativelikeness exists in *connections* among words. This is like human learning that "occurs through the modification of the brain's nervous connections" (Herculano-Houzel 2002: 102). In other words, learning lies in changes in connections between synapses (Owens and Tanner 2017). Those "changes" correspond to searching for the best parameters, a learning process in which deep learning finds the most optimal weights and biases to generalize input data. In this sense, just as how "learning occurs in connection" sounds abstract, nativelikeness is also such an elusive concept that it seems intangible to us.

The two examined factors, syntactic well-formedness and lexical associations, seem to be involved in determining nativelikeness to an extent, but they account only for some parts; nativelikeness needs to be considered with a wider view than the one here. As explaining nativelikeness only with the linguistic factors addressed here has inevitable limitations, future research involving a clearer investigation into nativelikeness could be conducted in two ways. First, we need to consider more subtypes of syntactic well-formedness and/or lexical associations; for example, more morphosyntactic and theta-role-based categories. Further, lexical associations can be subcategorized into semantic prosody (Stubbs 1995), semantic preference (Partington 2004), amplifier collocations (Altenberg 1991), etc. Second, it is necessary to consider factors beyond syntactic and lexical associations, such as phraseology and sociocultural contexts. Phraseological expressions are considered by many linguists as "one of the aspects that unmistakably distinguishes native speakers of a language from non-native speakers" (Ebeling and Hasselgård 2015: 185). Magnusson and Stroud (2012: 323), in their research on nativelikeness from the perspective of interactional sociolinguistics, note that "[l]anguage learners in late modern societies are heterogeneous and diverse," which means sociolinguistic approaches are needed to specify language learners' characteristics. In addition, shallow structure hypothesis (Clahsen and Felser 2006a), sentence length, word level, fluency, accuracy, automatization, and restructuring can also be considered.

Furthermore, in future research we expect to train the models with learner data made only by advanced English learners. In the current work, we used learner data made by all the proficiency levels of L2 learners to compare between learners and natives (as it is hard to collect enough advanced learner data to train deep neural networks). Using only near-nativelike sentences would facilitate a comparison between nativelikeness and near-nativelikeness.

6. Conclusion

The first research question of whether deep learning can sufficiently detect nativelikeness is substantiated by constructing binary classification models that have remarkably high accuracy. The models' ability to accurately discriminate between native and non-native data proves that they have learned nativelikeness, confirming that nativelikeness is measurable.

The second question concerns whether syntactic well-formedness and/or lexical associations are decisive factors influencing nativelikeness and how influential they are. The results are that neither is a determinant for judging nativelikeness. The heatmaps, except some that are not comprehensible, suggest that to some extent, well-formedness and lexical associations do affect the models' decisions but they do not seem that critical. Note that we do not argue that syntactic well-formedness and lexical associations have no impact on nativelikeness; the two attributes are not sufficient but necessary conditions for nativelikeness. Furthermore, the number of minimal pairs given the same predictions indicates that the models give predictions in an incorporated view. This is partially because they consider every token and connection among them, which is a characteristic of contextual embedding. In sum, we cannot fully explain nativelikeness with only two linguistic factors.

Although failing to provide a complete explanation of what nativelikeness is, this research has confirmed an important implication – that nativelikeness is measurable. The considerably high accuracy of the nativelikeness judgment models, especially BERT and XLNet, opens up the feasibility of quantitative and empirical linguistic research on nativelikeness via investigating deep learning's judgments.

This methodology exemplifies using AI technology as an implementing tool and linguistics as a normative role between deep learning and linguistics research on nativelikeness. Deep learning could serve as a good reference for defining and testing nativelikeness, and we expect it to be used for pedagogical applications such as resolving learnability problems arising from the vague notion of nativelikeness. Although future research efforts must be preceded by clear explanations of how deep learning reaches such a high accuracy in nativelikeness judgments to exclude the possibility of deep learning's superficial heuristic processing (McCoy et al. 2019), the models in this study can probably serve at least as a yardstick, allowing academic dialogues on nativelikeness to continue in future linguistics research.

We expect this research to contribute in three ways. First, with deep learning technologies, the current work proffers a new, supplementary (but not alternative) methodology to study nativelikeness, thus enabling us to attempt to probe the conditions of nativelikeness. This is beneficial for studying nativelikeness because it allows us to investigate the concept from another perspective and offers numerically comprehensive explanations. Second, this research describes whether syntactic well-formedness and lexical associations influence nativelikeness. Finally, this work broadens the scope of linguistics research using deep learning by implementing language models with learner data.

References

Abrahamsson, N. and K. Hyltenstam. 2009. Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning* 59(2), 249–306.

Alain, G. and Y. Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Altenberg, B. 1991. Amplifier collocations in spoken English. In S. Johansson and A. Stenström, eds., *English*

Computer Corpora: Selected Papers and Research Guide, 127–147. Berlin: Mouton de Gruyter.

Bengio, Y., I. Goodfellow and A. Courville. 2017. *Deep Learning* (Vol. 1). Massachusetts, USA: MIT press.

Birdsong, D. 2005. Nativelikeness and non-nativelikeness in L2A research. *International Review of pplied Linguistics in Language Teaching* 43(4), 319-328.

Bylund, E., N. Abrahamsson and K. Hyltenstam. 2012. Does first language maintenance hamper nativelikeness in a second language? A study of ultimate attainment in early bilinguals. *Studies in Second Language Acquisition* 34(2), 215-241.

Carlstrom, B. and N. Price. 2013. *Gachon Learner Corpus*. http://thegachonlearnercorpus.blogspot.kr

Clahsen, H. and C. Felser. 2006a. Continuity and shallow structures in language processing: A reply to our commentators. *Applied Psycholinguistics* 27(1), 107–126.

Clahsen, H. and C. Felser. 2006b. How native-like is non-native language processing? *Trends in Cognitive Sciences* 10(12), 564–570.

Davies, M. 2008. *The Corpus of Contemporary American English (COCA): 560 million words, 1990–present.* www.english-corpora.org/coca

Davies, M. 2013. Google Scholar and COCA-Academic: Two very different approaches to examining academic English. *Journal of English for Academic Purposes* 12(3), 155–165.

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

Ebeling, S. O. and H. Hasselgård. 2015. Learner corpora and phraseology. In S. Granger, G. Gilquin and F. Meunier, eds., *The Cambridge Handbook of Learner Corpus Research*, 207–230. Cambridge: Cambridge University Press.

Firth, J. R. 1961 [1957]. *Papers in Linguistics: 1934–1951*. London: Oxford University Press.

Gao, L. 2005. *Latin Squares in Experimental Design*. East Lansing: Michigan State University.

Goldberg, A. E. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press on Demand.

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

Goodfellow, I., Y. Bengio, A. Courville and Y. Bengio. 2016. *Deep Learning* (Vol. 1, No. 2). Cambridge: MIT press.

Graff, D., J. Kong, K. Chen and K. Maeda. 2003. *English Gigaword.* Philadelphia: Linguistic Data Consortium.

Guerssel, M., K. Hale, M. Laughren, B. Levin and J. W. Eagle. 1985. A cross-linguistic study of transitivity alternations. *Cls* 21(2), 48-63.

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

Herculano-Houzel, S. 2002. Do you know your brain? A survey on public neuroscience literacy at the closing of the decade of the brain. *The Neuroscientist* 8(2), 98–110.

Hochreiter, S. and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8), 1735–1780

Ivanova, I., M. J. Pickering, J. F. McLean, A. Costa and H. P. Branigan. 2012. How do people produce ungrammatical utterances? *Journal of Memory and Language* 67(3), 355-370.

Johnson, J. S. and E. L. Newport. 1989. Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology* 21(1), 60–99.

Ko, H., T. Ionin and K. Wexler 2010. The role of presuppositionality in the second language acquisition of

English articles. *Linguistic Inquiry* 41(2), 213–254.

Krohn, J., G. Beyleveld and A. Bassens. 2019. *Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence*. Addison-Wesley Professional.

Lambrecht, K. 1986. *Topic, Focus, and the Grammar of Spoken French*. Doctoral dissertation, Berkeley: University of California.

Leacock, C., M. Chodorow, M. Gamon and J. Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners: Second Edition*. Morgan & Claypool Publishers.

Linzen, T. 2019. What can linguistics and deep learning contribute to each other? Response to Pater. *Language* 95(1), e99–e108.

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

Long, M. H. 1990. Maturational constraints on language development. *Studies in Second Language Acquisition* 12(3), 251-285.

Magnusson, J. E. and C. Stroud. 2012. High proficiency in markets of performance: A sociocultural approach to nativelikeness. *Studies in Second Language Acquisition* 34(2), 321–345.

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

McCoy, R. T., E. Pavlick and T. Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*

Myers, J. 2009. The design and analysis of small-scale syntactic judgment experiments. *Lingua* 119(3), 425-444.

Nesselhauf, N. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics* 24(2), 223–242.

Owens, M. T. and K. D. Tanner. 2017. Teaching as brain changing: Exploring connections between neuroscience and innovative teaching. *CBE – Life Sciences Education* 16(2), fe2.

Paszke, A., S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga and A. Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS workshop*, 2017. California

Partington, A. 2004. "Utterly content in each other's company": Semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9(1), 131–156.

Pawley, A. and F. H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards and R. W. Schmidt, *Language and Communication,* 191–225. London: Longman.

Rhee, S. and C. Jung. 2012. Yonsei English Learner Corpus (YELC). In *Proceedings of the First Yonsei English Corpus Symposium*, 26–36. Seoul.

Rozovskaya, A. and D. Roth. 2010. Training paradigms for correcting errors in grammar and usage. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 154–162. Stroudsburg.

Rumelhart, D. E., G. E. Hinton and R. J. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323(6088), 533–536.

Schmitt, N. 1998. Quantifying word association responses: What is native-like? *System* 26(3), 389–401.

Shei, C. C. and H. Pain. 2000. An ESL writer's collocational aid. *Computer Assisted Language Learning* 13(2), 167–182.

Song, S. and E. Oh. 2017. In Defense of Comparisons between Formal and Informal Acceptability Judgments. *Studies in Generative Grammar* 27(4), 893-902.

Sorace, A. 2003. Near-nativelikeness. In C. J. Doughty and M. H. Long, eds., *The Handbook of Second*

Language Acquisition, 130–151. New York: Blackwell.

Sorace, A. and F. Keller. 2005. Gradience in linguistic data. *Lingua* 115(11), 1497-1524.

Sprouse, J., C. T. Schütze and D. Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua* 134, 219-248.

Stubbs, M. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language* 2(1), 23–55.

Tanaka, Y. 2001. Compilation of a multilingual corpus. In *Proceedings of the Pacific Association for Computational Linguistics (PACLING)*, 265–268. Kyushu.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Warstadt, A., A. Singh and S. R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics* 7, 625–641.

Warstadt, A., A. Parrish, H. Liu, A. Mohananey, W. Peng, S. Wang and S. R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics* 8, 377–392.

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

Wray, A. 2000. Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics* 21(4), 463–489.

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov and Q. V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*

Yoon, S., S. Park, J. Kim, H. Yoo and C. Jung. 2020. Incheon National University Multi-language Learner Corpus (INU-MULC): Its design and application. Abstract accepted at *Asia Pacific Corpus Linguistics Conference 2020*. Seoul, South Korea.

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

Appendix A

Test Suite II (DeKeyser 2000)

Type	Subtype	Example
Past tense	Past tense marking omitted in obligatory context	Sandy [filled/*fill] a jar with cookies last night.
	Irregular verbs regularized	Janie [slept/*sleeped] with her teddy bear last night.
	Regular ending on irregular stem	A bat [flew/*flewed] into our attic last night.
Plural	Plural marking omitted in obligatory context	The farmer bought two [pigs/*pig] at the market.
	Irregular plurals regularized	Two [mice/*mouses] ran into the house this morning.
	Mass nouns used with plural marker	Our neighbor bought new [furniture/*furnitures] last week.
Third-person singular	Third-person -s omitted in obligatory context	Every Friday our neighbor [washes/*wash] her car.
	Third-person -s marked on main verb after modals	John can [play/*plays] the piano very well.
Present progressive	Progressive -ing omitted in obligatory context	Janet is [wearing/*wear] the dress I gave her.
	Progressive auxiliary omitted	Tom [is working/*working] in his office right now.
Determiners	Determiner omitted in obligatory context	Mrs. Johnson went to [the library/*library] yesterday.
	Determiner used with abstract nouns	[Red/*The red] is a beautiful color.
Pronominalization	Pronoun omitted in obligatory context	Mike wrote the letter but didn't [send it/*send].
	Gender errors	John knew but [he/*she] did not tell.
Particle Movement	Phrasal verb separation not allowed	This plastic [gives off a weird smell/*gives a weird smell off].
	Phrasal verb separation allowed, but particle moved too far	She broke her shoes [in very carefully/*very carefully in].
Subcategorization	None	John [told/*said] me that his wife was ill.
Yes-No Questions	*aux Aux s[. . .	[Has the King been/*Has been the King] served his dinner?
	*aux Verb s[. . .	Can [the little girl ride/*ride the little girl] a bicycle?
	*V s[. . .	[Did Janet swim/*Swam Janet] in the race yesterday?
	Double tense marking	Where did Arnie [hunt/*hunted] last year?
Wh-Questions	No aux inversion	When [will Sam/*Sam will] fix his car?
	No aux	What [do they/*they] sell at the corner store?
Word Order	S V DO order violated	The [boy caught the ball/*ball the boy caught].
	S V IO DO order violated	The boy [feeds the rabbits carrots/*carrots feeds the rabbits].
	S V order violated	[The dog bites/*Bites the dog].
	S V PP order violated	The children [play with the dog/*with the dog play].
	Adverb placement	Kevin [usually rides/*rides usually] his bicycle to work.

Appendix B

Test Suite III

Test Set	Category	Example	Source
Well-formedness	Subject-verb agreement	My best friend [knows/*know] this guy.	(Leacock et al. 2010)
	Determiner	The marching band came past [the post office/*post office].	(Leacock et al. 2010)
	Negation	We [do not have/*have no] any time.	(Leacock et al. 2010)
	Argument	Most of the people went [to the beach/*the beach] on the weekend.	(Leacock et al. 2010)
	Reflexive pronoun	I want to make [myself/*me] fit.	(Leacock et al. 2010)
	Verb tense	I look forward to [seeing/*see] you.	(Leacock et al. 2010)
	Modal verb	People would [say/*said] that they want to play basketball.	(Leacock et al. 2010
	Infinitive clause	He is able to [begin/*began] a family.	(Leacock et al. 2010)
	Participle	Their parents are [expecting/*expect] good grades.	(Leacock et al. 2010)
	Predicative adjective	I am [interested/*interesting] in many things.	(Leacock et al. 2010)
	Auxiliary agreement	They have been [living/*live] here for 20 years.	(Leacock et al. 2010)
Lexical associations (collocations)	-	The company bought a [powerful/#strong] computer.	(Leacock et al. 2010)
		Tom [took/#ate] the pill.	(Leacock et al. 2010)
		They will [hold/#make] an election on September 9th.	(Leacock et al. 2010)
		Susan [conducted/#performed] a survey on the issue.	(Nesselhauf 2003)
		Mary wants to [shoot/#take] a film.	(Nesselhauf 2003)
		She should [do/#make] her homework.	(Nesselhauf 2003)
		The teacher [provided/#gave] a solution to the problem.	(Nesselhauf 2003)
		Mary must [carry out/#take] her own tasks.	(Shei and Pain 2000)
		Mary [transmitted/#conveyed] some important information.	(Shei and Pain 2000)
		They finally [reached/#made] an agreement.	(Shei and Pain 2000)
		Hard work [ensures/#warrants] success.	(Shei and Pain 2000)

Appendix C

Full versions of Test Suite II and Test Suite III/Full lists of Heatmaps for Test Suite II and Test Suite III

Website Link: https://bit.ly/2S5KqYO