Natural Language Processing Journal 6 (2024) 100060

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

Transformer-based text similarity and second language proficiency: A case of written production by learners of Korean

Gyu-Ho Shin [a,*], Boo Kyung Jung [b], Seongmin Mun [c]

[a] Department of Linguistics, University of Illinois Chicago, 601 S Morgan St, Chicago, IL 60607, USA
[b] Department of East Asian Languages & Literatures, University of Pittsburgh, 2714 Cathedral of Learning, Pittsburgh, PA 15260, USA
[c] Humanities Research Institute, Ajou University, 206 World cup-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, South Korea

ARTICLE INFO

Keywords:
Second language writing
Transformer
Similarity score
Proficiency
Human rating

ABSTRACT

The present study applies two transformer models (BERT; GPT-2) to analyse argumentative essays produced by two first-language groups (Czech; English) of second-language learners of Korean and investigates how informative similarity scores of learner writing obtained by these models explain general language proficiency in Korean. Results show three major aspects on model performance. First, the relationships between the similarity scores and the proficiency scores differ from the tendencies between the human rating scores and the proficiency scores. Second, the degree to which the similarity scores obtained by each model explain the proficiency scores is asymmetric and idiosyncratic. Third, the performance of the two models is affected by learners' native language and essay topic. These findings invite the need for researchers and educators to pay attention to how computational algorithms operate, together with learner language characteristics and language-specific properties of the target language, in utilising Natural Language Processing methods and techniques for their research or instructional purposes.

1. Introduction

With the recent advancement of Natural Language Processing (NLP) techniques, learner corpus research has progressed in revealing the developmental trajectories of second language (L2) learners by automatically analysing large-scale corpora of learner production (Meurers and Dickinson, 2017; Weiss and Meurers, 2021). Amongst many areas of NLP-assisted learner corpus research, *text quality* concerns the semantic–pragmatic aspects of language use in learner corpora (Burstein et al., 2013; Crossley and McNamara, 2013; Crossley et al., 2019; Cummins et al., 2016). Studies have shown positive relationships between the quality of writing and human raters' evaluation (Crossley and McNamara, 2013) and between the proficiency and similarity of spoken production to a test prompt (Crossley et al., 2019). As text quality is multifaceted and 'invisible' in nature, researchers often operationalise its measurement by employing concrete, 'visible' features such as coherence by way of cohesion devices and the degree of similarity relative to the native norm (Crossley et al., 2019), with the present study focusing on the latter. Particularly for the similarity feature, NLP techniques play a major role in data processing and measurement quantification (Burstein et al., 2013; Crossley et al., 2019; Dascalu et al., 2017; Panigrahi et al., 2018). To illustrate, Dascalu et al. (2017) investigated the semantic complexity of Dutch corpora with computational methodologies (e.g., *Wu-Palmer semantic distance, Latent Semantic Analysis*), showing meaningful correlations between text scores and the degree of paragraph elaboration measured by lexical diversity. Crossley et al. (2019) also conducted automatic text analysis for cohesion with semantic indices obtained from several NLP techniques (e.g., *Latent Semantic Analysis, Latent Dirichlet Allocation, Word2Vec*) and found considerable similarity between learners' oral production and test prompts in the TOEFL-iBT integrated speaking task as proficiency increased.

Despite the recent endeavour to automatically evaluate the text quality of learner corpora, we identify two major points for consideration. First, the roles of NLP techniques in text quality measurement need clarification concerning specific constructs of L2 competence. Many studies on text quality analysis have employed these techniques in various ways (Crossley and McNamara, 2013; Cummins et al., 2016; Panigrahi et al., 2018) but have not distinctively revealed how each technique addresses learner constructs such as proficiency. As these techniques differ in their assumptions and technical details of application, each technique may explain the same construct of interest differently. Hence, an investigation into how NLP techniques reveal learner constructs in text quality measurement is needed. Second, the application of NLP techniques to text quality assessment of learner corpora occurs in a restricted range of languages, primarily in L2-English. In contrast, few studies target L2s other than English. This sampling bias raises concern about whether and to what degree the

* Corresponding author.
E-mail address: ghshin@uic.edu (G.-H. Shin).

https://doi.org/10.1016/j.nlp.2024.100060
Received 1 December 2023; Received in revised form 3 February 2024; Accepted 18 February 2024

implications of previous studies hold for underrepresented languages, particularly those that differ typologically from the commonly sampled languages (cf. Bender et al., 2021).

This study aims to fill these gaps in two major directions. First, we employ L2 writing produced by learners of Korean. This language, understudied in the field, is characterised by the productive use of particles and suffixes attached to nominals and predicates, together with scrambling and omission of sentential components (Sohn, 1999). Its language-specific properties also make it a computationally challenging language (Kim et al., 2007; Shin and Jung, 2021). Despite the increasing popularity of L2-Korean worldwide, studies on the automatic analysis of L2-Korean learner corpora are lacking. Few studies have conducted text quality measurement (Cho and Park, 2018; Park and Lee, 2016), and they have grave issues regarding how they interpret their findings, control for learner background, and use NLP techniques. These issues limit the implications of these studies and weaken the reproducibility of procedures and results, which are essential to learner corpus research.

Second, we compare two computational models — Bidirectional Encoder Representations from Transformers (*BERT*; Devlin et al., 2018) and Generative Pre-trained Transformer 2 (*GPT-2*; Radford et al., 2019) — against the same dataset and learner construct. How neural networks capture human language behaviour is of growing interest amongst researchers (Contreras Kallens et al., 2023; Futrell and Levy, 2019; Oh et al., 2022; Warstadt et al., 2019; Xu et al., 2023; cf. Warstadt and Bowman, 2023 for a general overview); research has shown that the transformer architecture, which reduces sequential computation and relies entirely on an attention mechanism (Vaswani et al., 2017), yields better performance on language tasks than previously proposed recurrent models (Hawkins et al., 2020). Despite this significant finding, we are not aware of any empirical study that applies transformers to learner corpora, compares their performance in relation to learner constructs, or targets non-Indo-European language in this regard.

Together, we investigate the two models' performance concerning text similarity (as one specific and concrete index of text quality) with reference to general L2 proficiency (as a specific and concrete index of L2 competence). We specifically test how similarity scores of L2 writing (relative to native speaker writing) explain L2 proficiency (quantified through a separate measurement). Given the lack of studies that have pursued this inquiry in this manner, our study is innovative and explorative.

1.1. Two transformers: BERT and GPT-2

In computer science, a neural network, analogous to a biological neural network in a brain, is a computing system comprising weighted and layered interconnections between processing units (loosely modelling neurons in the brain) responding to input in parallel and producing output through propagation (see Kriesel, 2007 for in-depth descriptions of neural networks). After the initial proposal on neural-network computing in the 1940s, Rosenblatt (1958) developed a two-layer network, dubbed the *Perceptron*, which learns certain classifications by adjusting connection weights; this seminal work laid the foundations for later work in neural-network computing. Since then, the learning algorithms and procedures of neural-network systems have been steadily reconstructed, and they are now employed in various disciplines due to their efficient performance on data analysis factors (Wang et al., 2017). Through this development, neural networks in computing have become much less like biological neural networks (Crick, 1989), requiring a massive amount of training data for optimal operation (e.g., Edwards, 2015).

Transformer utilises the attention mechanism (Bahdanau et al., 2015) for effective computation. This mechanism identifies the most relevant parts of that sequence for processing by enhancing each part of the input sequence differently, considering various pieces of information about the whole sequence (see Vaswani et al., 2017 for in-depth descriptions of this architecture). This enhancement allows the transformer architecture to retain information from the early-appearing elements when handling long input sequences during information processing, thus achieving state-of-the-art performance in many downstream tasks (Ludwig et al., 2021; Vaswani et al., 2017).

In contrast to context-free word-embedding techniques, BERT (Devlin et al., 2018) proactively considers the context for each occurrence of a word through two specific procedures: *masked language model* (by masking some words in a sentence input randomly and predicting these masked words based only on the context suggested by other words surrounding the masked words) and *next sentence prediction* (by determining whether one sentence comes after another within the data in a binary manner) (see Devlin et al., 2018 for the technical details on BERT). The precise reasons for BERT's state-of-the-art performance on various downstream tasks remain under debate (Clark et al., 2019). However, an emerging line of research applies BERT to address linguistic inquiries (Hawkins et al., 2020; Warstadt and Bowman, 2020) and improves its performance by modifying model specifications. Some studies have employed BERT for automated essay scoring of L1 writing (see Ramesh and Sanampudi, 2022 for the extensive review on this work), demonstrating its strength and robustness in this task (Ludwig et al., 2021; Wangkriangkri et al., 2020) while showing its need for an exceedingly large number of parameters and its potential shortcomings in retaining previously learnt information from given datasets (Ormerod et al., 2021; Rodriguez et al., 2019). One recent study (Xue et al., 2021) extended this line of work by utilising BERT for scoring L1-Chinese L2-English learner writing, reporting a decent level of parity between BERT and human scorers.

GPT-2 (Radford et al., 2019) differs from BERT in motivation and internal composition. The former aims to accommodate a general-purpose learner whose learning trajectories are not subject to task types. Therefore, the model-training process does not rely on the specifics of data or tasks at hand. Moreover, fine-tuning through extensive modifications of hyperparameters or the architecture is not necessary, it learns what it must learn in an unsupervised (or self-supervised) manner. Such domain generality requires large datasets and many parameters from the outset. Indeed, GPT-2 has 1.5 billion parameters, trained on a dataset of eight million web pages, indicating the resource intensity in utilising this model (see Radford et al., 2019 for the technical details on GPT-2). Despite the continuous development of the GPT-*n* architecture, GPT-2 displays state-of-the-art performance in many language tasks (Goldstein et al., 2022a,b; Hosseini et al., 2022)), and with this enhanced capacity, researchers attempt to extend it to other lesser-studied languages (de Vries and Nissim, 2021). Liu et al. (2021), amongst others, investigated the parity between GPT-2 and human rating concerning various text quality indices (e.g., fluency, narrativity, language use). They compared the model's perplexity scores on several datasets, including Wikipedia articles and L1/L2 essays, all of which were written in English, with the human raters' evaluation scores. They found asymmetric degrees of correlation between the perplexity scores and the indices, also highlighting possible factors that may have affected the model performance: learner language characteristics, use of proper names or rare loanwords, genre-specific writing style, and so forth.

2. Methods

We developed two transformer models to analyse learner essays collected from two groups of learners whose L1s were either Czech (synthetic, highly inflectional, active agreement system, flexible word order) or English (analytic, little inflection, less active agreement system, fixed word order), which are typologically contrastive but both Subject–Verb–Object languages. We compared the relationship between proficiency and automatic measurement (similarity scores of learner writing, relative to native speaker writing, obtained from each model) and manual measurement (human rating scores) as a reference comparison. We note that our choices and decisions in conducting this

Table 1
Information about data by topic.

Topic	L2 learner						Native speaker		
	CZH			ENG			Mean (SD)	Min	Max
	Mean (SD)	Min	Max	Mean (SD)	Min	Max			
1	95.26 (33.18)	38	158	110.76 (29.82)	58	179	173.36 (57.85)	91	306
2	85.03 (30.26)	30	154	107.71 (35.96)	46	198	162.28 (52.48)	90	268
3	91.76 (32.42)	29	156	101.41 (32.34)	36	169	160.28 (52.41)	81	265

Note. The numeric values indicate the number of eojeols. An eojeol refers to a white-space-based segment which serves as a minimal language unit in Korean.

Table 2
Summary of model specification.

	BERT	GPT-2
Pre-trained model	*KoBERT*[a]	*KoGPT2-base-v2*[b]
Tokenisation	Syllable-based; *WordPiece*	Syllable-based; *Byte Pair Encoding*
Hyperparameters	Epoch: 30; Batch: 32; Sequence length: 256; Learning rate: .0001; Seed: 42; Epsilon: .00000001	

[a] https://github.com/SKTBrain/KoBERT
[b] https://github.com/SKT-AI/KoGPT2

study stand on resource-wise limitations when individual researchers in academia utilise large language models and employ computational procedures relating to those models (e.g., restricted access to the latest algorithms [e.g., GPT-4] and pre-trained models, weak computing power, high cost for using GPUs or external servers).

2.1. Data collection

We recruited a total of 68 learners from two groups: 34 L1-Czech L2-Korean (CZH; M_{age} = 24.0; SD = 2.69) and 34 L1-English L2-Korean (ENG; M_{age} = 26.1; SD = 4.43) learners, all non-heritage speakers of Korean and university students at the moment of testing. The amount of time spent in South Korea varied among participants (two to 10 years, M_{year} = 4.00, SD = 2.02 for CZH; one to 15 years, M_{year} = 5.68, SD = 3.36 for ENG). We also recruited 25 native speakers of Korean as a control group (M_{age} = 23.6, SD = 4.10). Participants joined an online meeting room and were asked to write three argumentative essays on a separate sheet of paper for 20 min each. We adapted essay topics from the Test of Proficiency in Korean (Topic 1: *Is early language education necessary for children?*; Topic 2: *Do we need to learn history?*; Topic 3: *Which do you prefer, competition or cooperation?*). The prompts were presented in Korean and each L1 to ensure the participants' clear understanding of these topics. The use of electronic devices was prohibited during the session.

Learner participants joined a proficiency measurement session through the Korean C-test (Lee-Ellis, 2009), comprising paragraphs with syllable-unit blanks, which test takers fill in based on each paragraph's context. The reliability and validity of this test in revealing general language proficiency were verified (Lee-Ellis, 2009; see also Eckes and Grotjahn, 2006; McKay, 2019). To improve the efficiency of the entire data collection process, we used the first four out of five excerpts in the test (the highest score: 188), following Lee-Ellis (2009). The participants' minimum and maximum scores from this test were 63 and 186, respectively (M_{score} = 103.74, SD = 25.66 for CZH; M_{score} = 110.59, SD = 30.30 for ENG). An independent samples t-test showed no statistical by-group difference in the scores ($t(66) = -1.006$, $p = .318$), indicating that proficiency in the two learner groups was not substantially different.

2.2. Data processing[1]

Two native Korean transcribers converted hand-written essays into machine-readable electronic files per participant and per topic, with typos and spelling/spacing errors retained (Table 1). We verified that their conversions were identical; if there was any mismatch, we reviewed all instances of disagreement and resolved them. Further inspection of the data revealed no direct use of the prompts in the essays.

Table 2 summarises the characteristics of the two transformer models that we developed. With the Python package *Transformers* and the pre-trained models respective to each model, we created our models by modifying model hyperparameters to obtain optimal outcomes, considering the recommendations and suggestions from previous studies (Alfaro et al., 2019; Dai et al., 2023; Kishimoto et al., 2020; de Vries and Nissim, 2021). No such pre-trained model for L2-Korean had been developed, and creating L2-Korean pre-trained models was impossible due to the limited data gathered, so we had to use the existing pre-trained models based on L1-Korean. We note that comparing variations of the transformer models with hyperparameter changes was not the primary interest of this study.

To compose input data, we created a data frame comprising rows indicating sentences of learner writing by topic and group and columns showing documents (including learner writing individually and native speaker writing as a whole) by group. This data frame was used to fine-tune each model: for BERT, every sentence in the rows contained [CLS] ('classification'; marking the start of a sentence) and [SEP] ('separation'; marking the end of a sentence) before and after each sentence, respectively, to indicate sentence boundaries; for GPT-2, no such addition occurred. The treatment for BERT was necessary due to the package that we used.

For model training, we first tokenised the sentences with the labels excluded. The maximum number of tokens in one sentence was set to 256 for the optimal model training; the model automatically trimmed any sentence that exceeded this limit. After that, we converted tokens in each sentence into 0 (*not attested*) or 1 (*attested*). All the information obtained by that process was transformed into a *tensor* – a data format reducing the size to increase processing speed. We then proceeded to the model-training process with a batch size of 32 for random sampling of the data per epoch to avoid excessive memory consumption. We parameterised the process in two ways: the epsilon had an initial value of .00000001 to prevent any division by zero, and the learning rate had an initial value of .0001. These values were automatically updated with the outcomes of each training epoch. The initial value allowing the model to run, *seed*, was set to 42. The training occurred 30 times per epoch with a batch size of 32, from the initial model with the zero value of gradients to an optimal model with updated values through forward- and back-propagation (cf. Xu et al., 2020). We chose the epoch size 30, with a minimised loss value (i.e., the difference between outcomes from a language model in a particular epoch and actual data) and constant afterwards. Finally, we obtained embedding outcomes, composing a

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

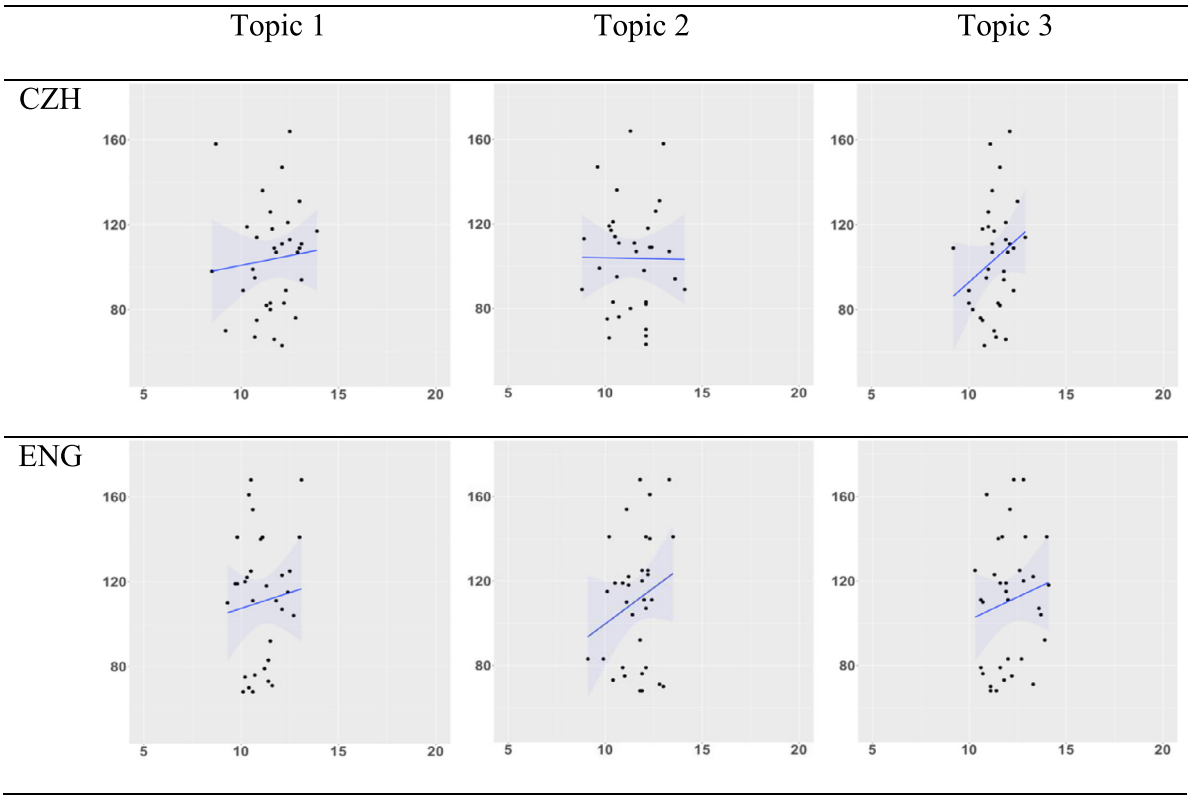

Fig. 1. Rating (averaged; *X*-axis) and proficiency (*Y*-axis) by L1 and topic. *Note*. The shaded areas indicate 95% CIs of each regression line.

total of sets (the number of sentences in the dataset) of arrays (the labels of writings in the dataset). To determine the representative value of each document, we excluded outliers from a cluster of values in each document produced by the model (cf. Breunig et al., 2000). Excluding outliers this way and using centre values in a cluster helped us to control for potential overfitting issues, thus affording us more reliable model outcomes. The trimmed data were reduced to two-dimensional embeddings using the t-SNE technique (Van der Maaten and Hinton, 2008).

We used these embeddings to calculate similarity scores between the centre value of the individual learners' writing and that of the native speakers' writing (as a whole) using cosine similarity as in Eq. (1).

$$cos(\theta) = \frac{A * B}{\|A\| \, \|B\|} = \frac{\sum_{i-1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} * \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

A_i and B_i are components of vectors A and B, respectively. $A*B$ is the dot product of the two vectors. $\|A\|$ and $\|B\|$ are the Euclidean norms of the two vectors (i.e., vector lengths), which are defined as $\sqrt{A_1^2 + A_2^2 + \cdots + A_i^2}$ and $\sqrt{B_1^2 + B_2^2 + \cdots + B_i^2}$, respectively. Cosine similarity was calculated by dividing the numerator by the denominator, producing values ranging from 0 (perfectly dissimilar) to 1 (perfectly similar).

2.3. Human rating

We recruited 10 raters (M_{age} = 32.9; SD = 8.69) who were experienced instructors of Korean (M_{year} = 3.5; SD = 3.48) and had relevant academic degrees. They were asked to evaluate learner essays for content, organisation, and language use (as employed in the writing evaluation process of the Test of Proficiency in Korean). Each category was measured on an eight-point Likert scale through an online platform (Table 3). We calculated each rater's score by summing all the scores from these three categories.

2.4. Statistical analysis

Similarity scores or human rating scores (independent variable) and proficiency scores (dependent variable) were fitted to three separate linear mixed-effects models (with *Rating* and *Group* as fixed effects and *Topic* and *Participant* as random effects for the rating–proficiency model; *Similarity* and *Group* as fixed effects and *Topic* and *Participant* as random effects for the BERT–proficiency and GPT-2–proficiency models) using the *lme4* software package (Bates et al., 2015) in *R* (R Core Team, 2023). Each model included the maximal random-effects structure allowed by the model (Barr et al., 2013). Considering the gradient nature of proficiency, we treated the proficiency scores as a continuous variable. We also computed each model's R^2 value using Nakagawa's R^2 (Nakagawa and Schielzeth, 2013; conditional R^2 considering both fixed and random effects). We note that we did not include time spent in Korea in the models as a random effect because of model convergence.

3. Results

3.1. Human rating and proficiency

Before evaluating the relationship between the similarity scores and the proficiency scores, we checked the relationship between the rating scores and the proficiency scores as a reference. Fig. 1 presents the relationships between human rating scores (averaged) and proficiency scores by learners' L1 and essay topics. We found that, except in the case of CZH's writing for T2, a positive tendency between human rating and proficiency was observed. However, the global model (α = .05) did not statistically support this observation (all ps > .05). Additional by-topic analyses (α = .025) revealed a main effect of *Rating* only for Topic 3 (β = 6.823, SE = 2.590, t = 2.635, p = .009, R^2: 0.306), indicating that the rating scores explained the proficiency scores only for this topic. We also conducted by-group analyses per topic through linear regression (due to the non-convergence of linear mixed-effects

Table 3
Human rating rubric.

Assessment type	Questions: "Did the writer…"
Content	perform the given task faithfully? compose the contents relevant to the topic? express the contents in a proper and diverse way?
Organisation	compose the contents clearly and logically? systematically connect the contents, using discourse markers that help the development of logic?
Language use	appropriately use grammar and vocabulary and in various/proper ways? use grammar, vocabulary, spelling, and so forth accurately? maintain formality considering the purpose and function of writing?

Table 4
Statistical outcome: Global model (α = .05).

	BERT (R^2 = 0.421)				GPT-2 (R^2 = 0.370)			
	β	*SE*	*t*	*p*	β	*SE*	*t*	*p*
(intercept)	102.885	3.279	31.375	< .001	103.372	3.355	30.816	< .001
Similarity	14.537	3.489	4.166	< .001	20.168	6.698	3.011	.003
Group	3.745	2.924	1.281	.201	4.452	3.997	1.114	.266
Similarity * Group	8.885	7.007	1.268	.206	8.126	13.725	0.592	.554

modelling; α = .0125) but revealed no significance in explaining the outcome variable (*Proficiency*) by the predictor variable (*Rating*). This result has several explanations, such as the relatively small data set and raters' conservatism for the rating (as shown by the small variability in the individual rating scores per topic). Nevertheless, human rating (obtained manually and holistically) seems to reveal learner constructs (proficiency in this study) to some degree and more or less consistently.

3.2. Similarity and proficiency

Fig. 2 presents the relationships between similarity scores and proficiency scores by model, learners' L1, and essay topic. Whereas the visual trend between the human rating scores and the proficiency scores was relatively uniform (although statistically insignificant), the similarity–proficiency relationships were visually idiosyncratic. This finding indicates that the similarity scores, obtained automatically from the transformer models, may fundamentally differ from the rating scores obtained holistically from human evaluation.

Tables 4 and 5 present the outcomes of the two statistical models (BERT–proficiency; GPT-2–proficiency). The global models revealed a main effect of *Similarity*, indicating that the similarity scores obtained from BERT and GPT-2 generally explained learner proficiency. However, additional by-group and by-topic analyses revealed a main effect of *Similarity* only for two cases: BERT, ENG, T3 and GPT-2, ENG, T2. This finding indicates that the extent to which the similarity scores from each transformer model explained the proficiency scores was contingent upon group and essay topic.

This inconclusive pattern of model performance bears two indications. One is that the operation of these models may have been asymmetrically influenced by such factors as learners' L1 and essay topics, thus generating eccentric performance in explaining proficiency. Previous studies have revealed the impact of learners' L1 (Ströbel et al., 2020) and essay topics (Yang and Kim, 2020) on L2 writing, but the inconsistencies in the by-model performance render it insufficiently clear to precisely evaluate their roles in addressing proficiency in this study. Hence, we acknowledge that the current results are not fully informative in revealing each model's sensitivity to these factors; more research is needed to address this issue. The other indication is that the models may not have been adept at extracting a centralised tendency from learner writing. We could not statistically confirm all the visual trends between the similarity scores and the proficiency scores. Moreover, there were large variances, and sometimes bipolarised classification (i.e., two big clusters around zero and one), of the essays' similarity scores, as demonstrated by the dispersion of individual data points in the scatterplots. These results imply that the models' capacity to capture a major trend explaining learner writing is limited, possibly resulting in the deviation between the similarity scores and the corresponding proficiency scores.

To address model performance against proficiency in more detail, we created two within-L1 proficiency groups with seven writings whose similarity scores were the highest or lowest in each model. Based on this grouping, we scrutinised two cases: one in which a participant was uniformly classified into either group by the two models for each topic (Table 6) the other in which a participant was classified into one group by one model but into the other group by the other model (Table 7). The results in Table 6 show that, although each model could consistently classify some participants' essays into the same proficiency groups, the extent to which this classification occurred varied by learners' L1 and essay topics. This finding indicates the variability in each model in computing the similarity scores, contributing to the inconsistency in the overall similarity–proficiency relationships. In particular, as Table 7 illustrates, the finding that the same participants could be classified into both proficiency groups depending upon the model used clearly indicates the by-model variability in handling the same data.

Although caution must be taken in interpreting these results due to the small sample size, they suggest that not all the models revealed learner constructs – proficiency – consistently and to the same extent.

4. General discussion

This study investigated the informativeness of text similarity computed by two transformer models using L2-Korean written production data, which was evaluated with reference to general proficiency in Korean. The findings revealed asymmetric degrees to which the similarity scores explained the proficiency scores, which largely differed from the rating–proficiency tendency. Examining how these models classified the essays into proficiency groups showed that the performance of these models varied by the learners' L1 and the essay topics.

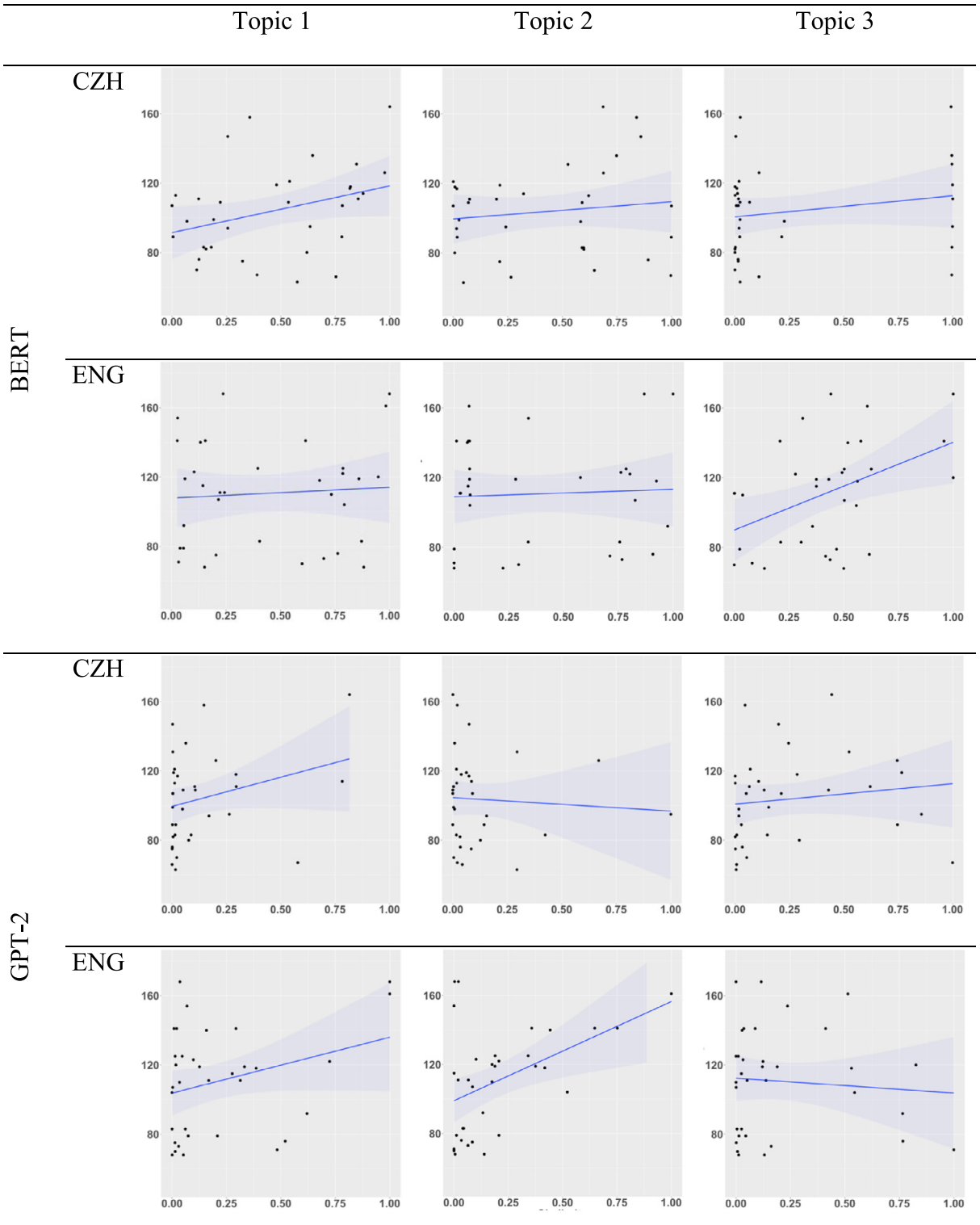

Fig. 2. Similarity (*X*-axis) and proficiency (*Y*-axis) by model, L1, and topic. *Note.* The shaded areas indicate 95% CIs of each regression line.

4.1. Why did the transformer models work this way?

We identify some promising factors that may have contributed to these outcomes. One factor lies in the learner language itself. Each model that we developed in this study was based on the representative pre-trained models created on large-scale and balanced L1 corpora, comprising typical L1 usage with varying styles and sophistication. In contrast, learner language involves considerable variability and ambiguity, including simple or short sentences and erroneous word use, often deviating from the target L1 usage (Meurers and Dickinson, 2017; O'Donnell et al., 2013). Furthermore, learners' use of proper names or rare words that would not have been found often in L1-Korean usage (e.g., *lwuiciayna* 'Louisiana', *hanmeli* 'one head [one person's idea]'), together with foreign words written in Korean characters (e.g., *monopolli*

Table 5
Statistical outcome: Additional analysis (α = .025).

			BERT				GPT-2			
			β	*SE*	*t*	*p*	β	*SE*	*t*	*p*
CZH	Topic 1	(intercept)	91.472	7.429	12.313	< .0005	99.546	5.007	19.882	< .0005
		Similarity	26.963	13.453	2.004	.054	33.670	20.679	1.628	.113
	Topic 2	(intercept)	99.613	6.972	14.288	< .0005	104.617	5.080	20.593	< .0005
		Similarity	9.867	12.887	0.766	.450	-7.778	21.450	-0.363	.719
	Topic 3	(intercept)	100.522	5.207	19.306	< .0005	100.861	5.796	17.402	< .0005
		Similarity	12.198	10.682	1.142	.262	11.759	15.294	0.769	.448
ENG	Topic 1	(intercept)	107.865	8.572	12.584	< .0005	103.752	6.373	16.280	< .0005
		Similarity	6.209	15.425	0.403	.690	32.258	18.401	1.753	.089
	Topic 2	(intercept)	108.995	7.566	14.406	< .0005	99.014	6.145	16.113	< .0005
		Similarity	4.224	14.389	0.294	.771	57.468	19.709	2.916	.006
	Topic 3	(intercept)	89.962	8.775	10.252	< .0005	112.328	6.516	17.238	< .0005
		Similarity	50.344	18.036	2.791	.009	-8.509	18.810	-0.452	.654

Note. We used linear regression for this analysis due to the non-convergence of the linear mixed-effects models.

Table 6
Participants uniformly classified into the same proficiency group for each topic.

		Highest	Lowest
CZH	Topic 1	6, 10, 14, 18, 23	22, 28
	Topic 2	5, 18	15, 21
	Topic 3	8, 9, 16, 33	11, 20, 25
ENG	Topic 1	19	–
	Topic 2	5	28, 33
	Topic 3	8, 19, 26	3, 18

'monopoly') may have additionally affected the similarity-score calculation process (cf. Liu et al., 2021). These aspects may have adversely affected the models' performance. Previous research has often argued that the direct application of NLP methods and techniques, which are trained exclusively on L1, general-purpose data, to L2 data may not yield reliable outcomes (Kyle, 2021; Meurers and Dickinson, 2017) because of the domain-specificity inherent in computational models. Relatedly, Sung and Shin (2023) showed L1 models' global failure to faithfully handle L2 datasets (and their improvements when fine-tuned on L2 training sets). Thus, further research is needed to assess whether and how fine-tuning currently available L1-based NLP tools on the target L2 data adjusts the tools' performance using these data.

Language-specific properties involving sentence formation in Korean, such as word order, case-marking, verbal morphology, and scrambling or omission of sentential components (Sohn, 1999), could also contribute to the unsatisfactory performance of these models (cf. Shin and Jung, 2021). To illustrate, L2 writers' infrequent or creative combinations of content nouns and case markers or unusual retention of sentential components, which would be absent in a native speaker's writing, would increase the unpredictability of subsequent tokens. This unpredictability may have aggravated these models' computations of learner writing relative to native speaker writing, leading them to assign similarity scores to the learner essays that deviated from the writer's proficiency score. Such deviations would have remained less prominent if we exclusively considered (L2-) English, which is structurally simpler and more straightforward than Korean. Accordingly, the scope of NLP-based research on learner corpora must be extended to under-represented/resourced languages to ensure a more nuanced understanding of its challenges and potential.

In addition to the two factors above, the characteristics of the transformer models' internal algorithms could also explain this discrepancy found in our results. The transformer architecture utilises raw sentences (with no Part-of-Speech information tagged) as a basic data-processing unit, as is typical for such modelling. Information about portions of the sentences obtained through tokenisation methods then predicts the items following these portions, assuming that a sequence comprises a context that allows it to share certain distributions or meanings (cf. distributional semantics hypothesis; Firth, 1957). Thus, the organisation of a sentence, or the collection of tokens in order, matters in the operation of these models. In this respect, the basic unit of data processing — sentence — may not be ideal for handling learner writing. In general, sentences that the learners produced were shorter in length than those from native speakers (Table 1). This simplicity, in addition to the characteristics of learner language and language-specific properties of the target language, may have inhibited these models from an ideal operation like what humans do in essay rating.

If our reasoning is reasonable and valid, it is broadly consistent with one general limitation to the current NLP techniques for text similarity (and beyond): they rely heavily on sequences of tokens — whether words, syllables, or characters. This characteristic renders it challenging for a computational model to identify a context involving semantic–pragmatic features (in a genuinely linguistic sense) during data processing, which is likely how humans evaluate text quality. Considering the transformer's state-of-the-art performance in many downstream tasks, future research should clarify whether and to what degree the algorithms of transformer models access and reveal

Table 7
Participants differently classified into the proficiency groups for each topic.

	CZH			ENG		
	Participant	BERT	GPT-2	Participant	BERT	GPT-2
Topic 1				9	Highest	Lowest
				12	Highest	Lowest
				33	Lowest	Highest
Topic 2	23	Highest	Lowest	34	Highest	Lowest
	17	Lowest	Highest	14	Lowest	Highest
Topic 3				6	Highest	Lowest
				33	Lowest	Highest

contextual/discourse information from the given text in automatically evaluating (learner) writing.

To clarify, we are not claiming that these models do not perform well with all learner corpora; their performance with other learner corpora should be addressed in more detail by subsequent research with various (and larger) L2 data. Moreover, comparing multiple sub-models with the same architecture(s) but hyperparameter variations is not the focus of the current study. A line of research has revealed the asymmetric performance of a computational model on addressing human language behaviour contingent upon architecture type or manipulation of the model's hyperparameters (Hu et al., 2020; Shin and Mun, 2023). In this regard, the findings of this study need to be re-assessed in terms of architectures and model hyperparameters, which is an important avenue for future research.

4.2. Broader implications in the field

Given the recent trend of NLP techniques being widely used in learner corpus research, this study's findings suggest that utilising computational models for research or instruction must be grounded on a sound understanding of how algorithms work and the various factors that could affect their operation. Our findings suggest that language educators must be knowledgeable about the selection and application of these models to learner corpora. The application of computational technology, represented as artificial intelligence (AI), to educational fields has recently accelerated due to rapid environmental changes, such as the COVID-19 pandemic and the high demand for individualised support for students and classrooms (Pai et al., 2020; Toncic, 2020). However, within the AI-to-education context, a gap exists between *deploying* AI and *understanding* AI: educators' competence in critically evaluating computational technology and properly utilising it appears not to keep abreast with current understanding (Ottenbreit-Leftwich et al., 2010; Tondeur et al., 2012). While some researchers actively offer the research–technology interface (Crossley et al., 2019; Kyle, 2016; Lu, 2010), the majority in this field remain consumers of the existing tools. Merely learning how to use AI-based tools and not understanding how they work tends to cause practitioners to struggle in real-world education settings (Bullock, 2016), rendering their endeavours incomplete.

Therefore, enhancing language educators' understanding of computational technology may be essential to improving its application in educational contexts while critically evaluating its strengths and weaknesses (Holmes et al., 2019; Seufert et al., 2021). This skill, dubbed *AI literacy*, involves basic knowledge of algorithms – central to AI (Holmes et al., 2019) – and basic programming skills that enable one to answer questions such as "How similarly or differently do various architectures operate?" "In what aspect do model hyperparameters relate to outcomes for data processing?" and "To what extent can we explain the results by algorithms?" An emerging line of research supports this view (Long and Magerko, 2020; Ng et al., 2021), suggesting that providing educators with the fundamentals of computational technology during teacher training will improve AI literacy and promote active communication with computer scientists and NLP specialists in dealing with technology-centred issues in education. The implications of our results closely align with this perspective: automatically computed text similarity varies by model, and without knowledge about the transformer architecture, one cannot interpret the outcomes properly. AI literacy can alleviate this issue, ensuring practitioners understand why this variation happens and can make informed decisions about which techniques to choose based on their research or educational needs.

5. Conclusion

Although the current study generated meaningful findings, it has its own limitations. First, we considered text similarity and proficiency as a proxy for text quality and learner constructs, respectively. However, text quality and learner constructs are multifaceted concepts, so relying only on two specific aspects may limit the implications of this study. More indices of text quality (e.g., discourse coherence), together with various learner constructs, must be considered to elucidate the precise relationship between these factors. Second, the input that our transformer models encountered — a sentence with no correction of typos and spelling or spacing errors — may have led them to be somewhat oblivious to properly measuring inter-sentential elaboration of ideas. Thus, although challenging, incorporating paragraph-level ideas, with different approaches to handling sentences, into the automatic evaluation of learner writing will benefit researchers.

Third, our choices and decisions in executing these simulations were based on the limitations on (i) utilising large language models and their related computational procedures in an ideal way (e.g., restricted access to the latest algorithms and pre-trained models, weak computing power, high cost for using GPUs or external servers) and (ii) working with small-scale learner corpora (due to the manual processes involving data collection [essay, proficiency] and essay grading [by human raters]). Although we tried our best to acquire and secure the resources required to conduct the simulations at the moment of study, our endeavour was notably restricted in various aspects. While we believe that the implications of this study still support the possible factors that we identified pertaining to addressing model performance and the necessity of AI literacy, the resource-wise limitations in the current study could prevent us from fully justifying the two models' behaviours in the given simulation environments. Therefore, replicating this study with enhanced computational resources and a larger sample size would be required to verify its conclusions, especially concerning the unsatisfactory performance of the models. This is what we plan to do next.

These limitations notwithstanding, the findings of the present study shed light on how informative transformers are of learner constructs. Based on the findings, researchers should consider the algorithmic characteristics of these models, together with learner- and target-language properties, in the automatic processing of learner corpora.

CRediT authorship contribution statement

Gyu-Ho Shin: Writing – review & editing Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Boo Kyung Jung:** Writing – review & editing Resources, Investigation, Conceptualization. **Seongmin Mun:** Writing – review & editing Visualization, Software, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Seed Program for Korean Studies through the Ministry of Education of the Republic of Korea and Korean Studies Promotion Service of the Academy of Korean Studies (AKS-2022-INC-2250004).

References

Alfaro, Felipe, Costa-jussà, Marta R., Fonollosa, José, 2019. Bert masked language modeling for coreference resolution. In: Costa-jussà, Marta R., Hardmeier, Christian, Radford, Will, Webster, Kellie (Eds.), Proceedings of the First Workshop on Gender Bias in Natural Language Processing. Association for Computational Linguistics, pp. 76–81.

Bahdanau, Dzmitry, Cho, Kyunghyun, Bengio, Yoshua, 2015. Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations. http://dx.doi.org/10.48550/arXiv.1409.0473.

Barr, Dale J., Levy, Roger, Scheepers, Christoph, Tily, Harry J., 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. J. Memory Lang. 68 (3), 255–278.

Bates, Douglas, Mächler, Martin, Bolker, Ben, Walker, Steve, 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1), 1–48.

Bender, Emily M., Gebru, Timnit, McMillan-Major, Angelina, Shmitchell, Shmargaret, 2021. On the dangers of stochastic parrots: Can language models be too big?. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 610–623.

Breunig, Markus M., Kriegel, Hans-Peter, Ng, Raymond T., Sander, Jörg, 2000. LOF: Identifying density-based local outliers. Sigmod Rec. 29 (2), 93–104.

Bullock, Shawn, 2016. Digital technologies in teacher education: From mythologies to making. In: Kosnik, Clare, White, Simone, Beck, Clive, Marshall, Bethan, Goodwin, Lin, Murray, Jean (Eds.), Building bridges: Rethinking Literacy Teacher Education in a Digital Era. Sense Publishers, Rotterdam, the Netherlands, pp. 3–16.

Burstein, Jill, Tetreault, Joel, Chodorow, Martin, Blanchard, Daniel, Andreyev, Slava, 2013. Automated evaluation of discourse coherence quality in essay writing. In: Shermis, Mark, Burstein, Jill (Eds.), Handbook of Automated Essay Evaluation: Current Applications and New Directions. Routledge, New York, pp. 267–280.

Cho, Sukyeon, Park, Youngmin, 2018. Sheffield tayhakkyo hankwuke haksupcauy cakmwun thukseng pwunsek [characteristics of Korean language writing by students at the university of sheffield]. Cakmwunyeonkwu [Writing Res.] 38, 149–172.

Clark, Kevin, Khandelwa, Urvashi, Levy, Omer, Mannin, Christopher, 2019. What does BERT look at? An analysis of bert's attention. In: Tal Linzen, Yonatan Belinkov & Dieuwke Hupkes (Ed.), Proceedings of the 2019 Association for Computational Linguistics Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, pp. 276–286.

Contreras Kallens, Pablo, Kristensen-McLachlan, Ross Deans, Christiansen, Morten H., 2023. Large language models demonstrate the potential of statistical learning in language. Cogn. Sci. 47 (3), e13256.

Crick, Francis, 1989. The recent excitement about neural networks. Nature 337 (6203), 129–132.

Crossley, Scott, Kyle, Kristopher, Dascalu, Mihai, 2019. The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. Behav. Res. Methods 51, 14–27.

Crossley, Scott, McNamara, Danielle, 2013. Applications of text analysis tools for spoken response grading. Lang. Learn. Technol. 17 (2), 171–192.

Cummins, Ronan, Yannakoudakis, Helen, Briscoe, Ted, 2016. Unsupervised modeling of topical relevance in L2 learner text. In: Tetreault, Joel, Burstein, Jill, Leacock, Claudia, Yannakoudakis, Helen (Eds.), Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics, pp. 95–104.

Dai, Damai, Sun, Yutao, Dong, Li, Hao, Yaru, Ma, Shuming, Sui, Zhifang, Wei, Furu, 2023. Why can GPT learn in-context? Language models secretly perform gradient descent as meta-optimizers. In: A. Rogers, & N. Okazaki (Ed.), Findings of the Association for Computational Linguistics. ACL 2023, Association for Computational Linguistics, pp. 4005–4019.

Dascalu, Mihai, Westera, Wim, Ruseti, Stefan, Trausan-Matu, Stefan, Kurvers, Hub, 2017. ReaderBench learns building a comprehensive automated essay scoring system for dutch language. In: André, Elizabeth, Baker, Ryan, Hu, Xiangen, Rodrigo, Ma. Mercedes, du Boulay, Benedict (Eds.), Artificial Intelligence in Education 2017. In: Lecture Notes in Computer Science, vol. 10331, Springer, pp. 52–63.

de Vries, Wietse, Nissim, Malvina, 2021. As good as new. How to successfully recycle english GPT-2 to make models for other languages. In: Zong, Chengqing, Xia, Fei, Li, Wenjie, Navigli, Roberto (Eds.), Findings of the Association for Computational Linguistics 2021. Association for Computational Linguistics, pp. 836–846.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, Toutanova, Kristina, 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, Jill, Doran, Christy, Solorio, Thamar (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, pp. 4171–4186.

Eckes, Thomas, Grotjahn, Rüdiger, 2006. A closer look at the construct validity of C-tests. Lang. Test. 23 (3), 290–325.

Edwards, Chris, 2015. Growing pains for deep learning. Commun. ACM 58 (7), 14–6.

Firth, John R., 1957. A synopsis of linguistic theory, 1930–1955. Stud. Linguist. Anal. 1–32.

Futrell, Richard, Levy, Roger P., 2019. Do RNNs learn human-like abstract word order preferences? In: Jarosz, Gaja, Nelson, Max, O'Connor, Brendan, Pater, Joe (Eds.), Proceedings of the Society for Computation in Linguistics 2019. pp. 50–59.

Goldstein, Ariel, Ham, Eric, Nastase, Samuel A., Zada, Zaid, Grinstein-Dabus, Avigail, Aubrey, Bobbi, Schain, Mariano, Gazula, Harshvardhan, Feder, Amir, Doyle, Werner, Devore, Sasha, Dugan, Patricia, Friedman, Daniel, Brenner, Michael, Hassidim, Avinatan, Devinsky, Orrin, Flinker, Adeen, Levy, Omer, Hasson, Uri, 2022b. Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain. http://dx.doi.org/10.1101/2022.07.11.499562, BioRxiv.

Goldstein, Ariel, Zada, Zaid, Buchnik, Eliav, Schain, Mariano, Price, Amy, Aubrey, Bobbi, Nastase, Samuel A., Feder, Amir, Emanuel, Dotan, Cohen, Alon, Jansen, Aren, Gazula, Harshvardhan, Choe, Gina, Rao, Aditi, Kim, Catherine, Casto, Colton, Fanda, Lora, Doyle, Werner, Friedman, Daniel, Dugan, Patricia, Melloni, Lucia, Reichart, Roi, Devore, Sasha, Flinker, Adeen, Hasenfratz, Liat, Levy, Omer, Hassidim, Avinatan, Brenner, Michael, Matias, Yossi, Norman, Kenneth A., Devinsky, Orrin, Hasson, Uri, 2022a. Shared computational principles for language processing in humans and deep language models. Nature Neurosci. 25, 369–380.

Hawkins, Robert D., Yamakoshi, Takateru, Griffiths, Thomas L., Goldberg, Adele E., 2020. Investigating representations of verb bias in neural language models. In: Webber, Bonnie, Cohn, Trevor, He, Yulan, Liu, Yang (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 4653–4663.

Holmes, Wayne, Bialik, Maya, Fadel, Charles, 2019. Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Center for Curriculum Redesign, Boston, MA.

Hosseini, Eghbal A., Schrimpf, Martin, Zhang, Yian, Bowman, Samuel, Zaslavsky, Noga, Fedorenko, Evelina, 2022. Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. BioRxiv, 2022-10.

Hu, Jennifer, Gauthier, Jon, Qian, Peng, Wilcox, Ethan, Levy, Roger P., 2020. A systematic assessment of syntactic generalization in neural language models. In: Jurafsky, Dan, Chai, Joyce, Schluter, Natalie, Tetreault, Joel (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 1725–1744.

Kim, Byoung-Soo, Lee, Yong-Hun, Lee, Jong-Hyeok, 2007. Unsupervised semantic role labeling for Korean adverbial case. J. KIISE: Softw. Appl. 34 (2), 32–39.

Kishimoto, Yudai, Murawaki, Yugo, Kurohashi, Sadao, 2020. Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In: Calzolari, Nicoletta, Béchet, Frédéric, Blache, Philippe, Choukri, Khalid, Cieri, Christopher, Declerck, Thierry, Goggi, Sara, Isahara, Hitoshi, Maegaard, Bente, Mariani, Joseph, Mazo, Hélène, Moreno, Asuncion, Odijk, Jan, Piperidis, Stelios (Eds.), Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, pp. 1152–1158.

Kriesel, David, 2007. A brief introduction to neural networks. Retrieved at http://www.dkriesel.com on 15-October-2020.

Kyle, Kristopher, 2016. Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication. Unpublished doctoral Dissertation. Georgia State University.

Kyle, Kristopher, 2021. Natural language processing for learner corpus research. Int. J. Learner Corpus Res. 7 (1), 1–16.

Lee-Ellis, Sunyoung, 2009. The development and validation of a Korean C-test using rasch analysis. Lang. Test. 26 (2), 245–274.

Liu, Yang, Medlar, Alan, Glowacka, Dorota, 2021. Can language models identify wikipedia articles with readability and style issues? In: Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval. Association for Computing Machinery, New York, NY, pp. 113–117.

Long, Duri, Magerko, Brian, 2020. What is AI literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, pp. 1–16.

Lu, X., 2010. Automatic analysis of syntactic complexity in second language writing. Internat. J. Corpus Linguistics 15 (4), 474–496.

Ludwig, Sabrian, Mayer, Christian, Hanse, Christopher, Eilers, Kerstin, Brandt, Steffen, 2021. Automated essay scoring using transformer models. Psych 3 (4), 897–915.

McKay, Todd, 2019. More on the Validity and Reliability of C-Test Scores: A Meta-Analysis of C-Test Studies (Ph.D. Dissertation). Georgetown University, Unpublished.

Meurers, Detmar, Dickinson, Markus, 2017. Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. Lang. Learn. 67 (S1), 66–95.

Nakagawa, Shinichi, Schielzeth, Holger, 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. Methods Ecol. Evol. 4 (2), 133–142.

Ng, Davy Tsz Kit, Leung, Jac Ka Lok, Samuel, Kai Wah, Qiao, Maggie Shen, 2021. AI literacy: Definition, teaching, evaluation and ethical issues. Proc. Assoc. Inf. Sci. Technol. 58 (1), 504–509.

O'Donnell, Matthew B., Römer, Ute, Ellis, Nick C., 2013. The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. Int. J. Corpus Linguist. 18 (1), 83–108.

Oh, Byung-Doh, Clark, Christian, Schuler, William, 2022. Comparison of structural parsers and neural language models as surprisal estimators. Front. Artif. Intell. 5, 777963.

Ormerod, Christopher M., Malhotra, Akanksha, Jafari, Amir, 2021. Automated essay scoring using efficient transformer-based language models. https://arxiv.org/abs/2102.13136.

Ottenbreit-Leftwich, Anne, Glazewski, Krista D., Newby, Timothy J., Ertmer, Peggy A., 2010. Teacher value beliefs associated with using technology: addressing professional and student needs. Comput. Educ. 55, 1321–1335.

Pai, Kai-Chih, Kuo, Bor-Chen, Liao, Chen-Huei, Liu, Yin-Mei, 2020. An application of Chinese dialogue-based intelligent tutoring system in remedial instruction for mathematics learning. Educ. Psychol. 41 (2), 137–152.

Panigrahi, Sabitra S., Panigrahi, Narayan, Paul, Biswajit, 2018. Modelling of topic from hindi corpus using Word2Vec. In: Proceedings of the 2018 Second International Conference on Advances in Computing, Control and Communication Technology. Institute of Electrical and Electronics Engineers, pp. 97–100.

Park, Jungyeul, Lee, Jung Hee, 2016. A Korean learner corpus and its features. Enehak [Linguist. Soc. Korea] 75, 69–85.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Radford, Alec, Wu, Jeffrey, Child, Rewon, Luan, David, Amodei, Dario, Sutskever, Ilya, 2019. Language models are unsupervised multitask learners. OpenAI Blog 1 (8), 9.

Ramesh, Dadi, Sanampudi, Suresh Kumar, 2022. An automated essay scoring systems: A systematic literature review. Artif. Intell. Rev. 55, 2495–2527.

Rodriguez, Pedro Uria, Jafari, Amir, Ormerod, Christopher M., 2019. Language models and automated essay scoring. https://arxiv.org/abs/1909.09482v1.

Rosenblatt, Frank, 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychol. Rev. 65 (6), 386–408.

Seufert, Sabine, Guggemos, Josef, Sailer, Michael, 2021. Technology-related knowledge, skills, and attitudes of pre-and in-service teachers: the current situation and emerging trends. Comput. Hum. Behav. 115, 106552.

Shin, Gyu-Ho, Jung, Boo Kyung, 2021. Automatic analysis of learner corpora in Korean: Written production of Korean passive constructions for mandarin-speaking learners of Korean. Int. J. Learner Corpus Res. 7 (1), 53–82.

Shin, Gyu-Ho, Mun, Seongmin, 2023. Explainability of neural networks for child language: Agent-first strategy in comprehension of Korean active transitive construction. Dev. Sci. e13405.

Sohn, Ho-Min, 1999. The Korean Language. Cambridge University Press, Cambridge, NY.

Ströbel, Marcus, Kerz, Elma, Wiechmann, Daniel, 2020. The relationship between first and second language writing: Investigating the effects of first language complexity on second language complexity in advanced stages of learning. Lang. Learn. 70 (3), 732–767.

Sung, Hakyung, Shin, Gyu-Ho, 2023. Diversifying language models for lesser-studied languages and language-usage contexts: A case of second language Korean. In: The Findings of EMNLP 2023.

Toncic, Jason, 2020. Teachers, AI grammar checkers, and the newest literacies: emending writing pedagogy and assessment. Digital Cult. Educ. 12 (1), 26–51.

Tondeur, Jo, van Braak, Johan, Sang, Guoyuan, Voogt, Joke, Fisser, Petra, Ottenbreit-Leftwich, Anne, 2012. Preparing preservice teachers to integrate technology in education: a synthesis of qualitative evidence. Comput. Educ. 59, 134–144.

Van der Maaten, Laurens, Hinton, Geoffrey, 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Łukasz, Polosukhin, Illia, 2017. Attention is all you need. In: von Luxburg, Ulrike, Guyon, Isabelle, Bengio, Samy, Wallach, Hanna, Fergus, Rob, Vishwanathan, S.V.N., Garnett, Ron (Eds.), Proceedings of the 31st Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 5998–6008.

Wang, Ding, He, Haibo, Liu, Derong, 2017. Intelligent optimal control with critic learning for a nonlinear overhead crane system. IEEE Trans. Ind. Inform. 14 (7), 2932–2940.

Wangkriangkri, Phakawat, Viboonlarp, Chanissara, Rutherford, Attapol T., Chuangsuwanich, Ekapol, 2020. A comparative study of pretrained language models for automated essay scoring with adversarial inputs. In: 2020 IEEE Region 10 Conference. IEEE, pp. 875–880.

Warstadt, Alex, Bowman, Samuel R., 2020. Can neural networks acquire a structural bias from raw linguistic data? In: Denison, Stephanie, Mack, Michael, Xu, Yang, Armstrong, Blare (Eds.), Proceedings of the 42nd Annual Conference of the Cognitive Science Society. Cognitive Science Society, pp. 1737–1743.

Warstadt, Alex, Bowman, Samuel R., 2023. What artificial neural networks can tell us about human language acquisition. In: Lappin, Shalom, Bernardy, Jean-Philippe (Eds.), Algebraic Structures in Natural Language. Taylor & Francis, Boca Raton, FL, pp. 17–60.

Warstadt, Alex, Singh, Amanpreet, Bowman, Samuel R., 2019. Neural network acceptability judgments. Trans. Assoc. Comput. Linguist. 7, 625–641.

Weiss, Zarah, Meurers, Detmar, 2021. Analyzing the linguistic complexity of German learner language in a reading comprehension task: Using proficiency classification to investigate short answer data, cross-data generalizability, and the impact of linguistic analysis quality. Int. J. Learner Corpus Res. 7 (1), 83–130.

Xu, Weijie, Chon, Jason, Liu, Tianran, Futrell, Richard, 2023. The linearity of the effect of surprisal on reading times across languages. In: Findings of the Association for Computational Linguistics. EMNLP 2023, pp. 15711–15721.

Xu, Ying, Qiu, Xipeng, Zhou, Ligao, Huang, Xuanjing, 2020. Improving BERT fine-tuning via self-ensemble and self-distillation. J. Comput. Sci. Tech. 33 (1), 1–18.

Xue, Jin, Tang, Xiaoyi, Zheng, Liyan, 2021. A hierarchical BERT-based transfer learning approach for multi-dimensional essay scoring. IEEE Access 9, 125403–125415.

Yang, Weiwei, Kim, YouJin, 2020. The effect of topic familiarity on the complexity, accuracy, and fluency of second language writing. Appl. Linguist. Rev. 11 (1), 79–108.

Gyu-Ho Shin, Ph.D. is an assistant professor in the Department of Linguistics at the University of Illinois at Chicago. His research interests include language acquisition/development, psychology of language, computational linguistics, and corpus linguistics. He has published articles in international journals such as *Applied Linguistic Review, Cognitive Development, Cognitive Science, Developmental Science, International Journal of Learner Corpus Research*, *Journal of Child Language*, and *Studies in Second Language Acquisition*.

Boo Kyung Jung, Ph.D. is an instructor of Korean in the Department of East Asian Languages and Literatures at the University of Pittsburgh. Her research interests include Korean pedagogy, corpus linguistics, and L2 writing. She has published articles in international journals such as *Australian Review of Applied Linguistics, Corpora*, and *International Review of Applied Linguistics in Language Teaching*.

Seongmin Mun, Ph.D. is a research assistant professor in the Humanities Research Institute at Ajou University. His expertise includes machine learning, natural language processing, and data visualisation. He has published articles in international journals such as *Applied Sciences, Infancy,* and *Language and Information*.