To what extent neural network models reveal L2 constructs? Relationship between text similarity and learner proficiency

With the recent development of NLP techniques, a number of second language (L2) studies utilise these techniques to automatically analyse learner corpora (Meurers, 2015). One area in learner corpus research is text quality, which concerns semantic–pragmatic aspects of language use to influence overall text quality (e.g., Crossley et al., 2019). Despite increasing interests in employing various NLP techniques (e.g., Dascalu et al., 2017), little attention has been paid to how similarly/differently each technique reveals L2 constructs such as learner proficiency. In addition, NLP-based L2 research is heavily biased towards L2-English, which does not ensure the generalisability of its implications. Against this background, we investigate the relationship between learner proficiency and text similarity of L2-Korean learners' written production (relative to native speakers' writing) measured through neural network models.

Method (Table 1). Thirty-six L1-Chinese L2-Korean learners (age: mean = 24.2; *SD* = 3.11) were asked to write argumentative essays on two topics: *preservation vs. exploitation of the nature; competition vs. cooperation*. Learner proficiency was measured separately, using the Korean C-test (Lee-Ellis, 2009; ranging from 0 to 188; mean = 135.98; *SD* = 32.23). Essays from 10 native Korean speakers were collected as a reference text. After extracting content words from the essays, we computed cosine similarity scores between individual learner writing and the reference text by employing two neural network models—Word2Vec (Mikolov et al., 2013; bag-of-words; context-independent) and BERT (Devlin et al., 2018; transformer; context-dependent). The similarity scores (predictor) and proficiency scores (outcome) were then submitted to linear regression models.

Results (Figure 1 & Table 2). The Word2Vec model showed a significant relationship between the two variables for both topics: $F(1, 34) = 3.405$, $p = .074$, $R^2 = .064$, $B = 113.86$ for Topic 1 (albeit marginal); $F(1, 34) = 8.748$, $p = .006$, $R^2 = .181$, $B = 172.59$ for Topic 2. For the BERT model, the slope of the regression line for Topic 1 was nearly horizontal whereas that for Topic 2 was positively oblique. This was reflected in the linear regression analysis, with marginal significance only for Topic 2: $F(1, 34) = 3.79$, $p = .060$, $R^2 = .074$, $B = 32.296$. To further examine how each neural network model classified the participants into the same group uniformly, we created two proficiency groups (highest; lowest) with seven essays by topic. These models demonstrated distinctive classification patterns, yielding weak congruency across the topics/models.

Together, these results indicate that (i) the degree that neural network models explain L2 constructs (learner proficiency in this study) was asymmetric and (ii) these models' performance was sensitive to essay topics (and particularly to word use such as repetitions of keywords), manifesting some limits on addressing individual variability of L2 writing as well. Given the recent trend that NLP techniques are widely used in learner corpus research, our findings suggest the need for researchers to be aware of NN models' algorithmic characteristics, together with possible influences of topic variations, in conducting automatic L2 text analysis research in pursuit of addressing L2 constructs.

Table 1. Information about data by topic (numeric values = number of words)

Topic	L2 learner			Native speaker		
	Mean (SD)	Minimum	Maximum	Mean (SD)	Minimum	Maximum
1	107 (36.36)	62	201	158 (21.27)	131	194
2	113 (38.48)	57	203	166 (33.89)	110	211

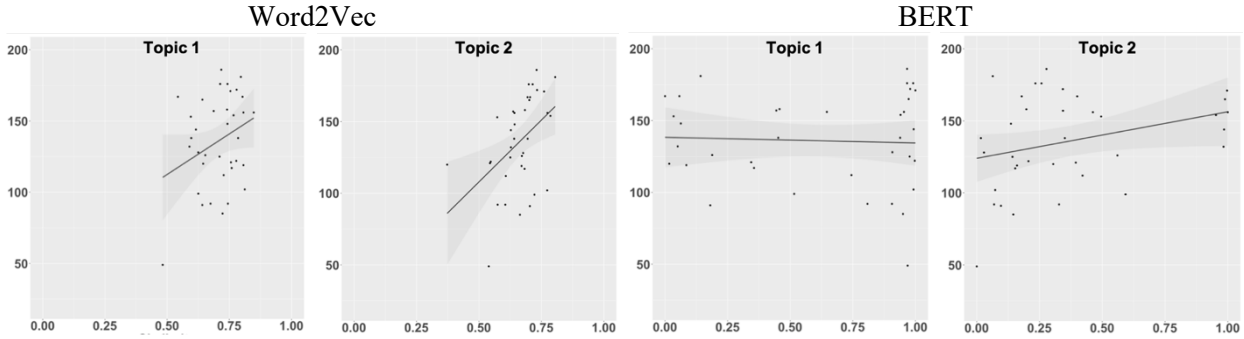

Figure 1. Scatterplot: Similarity scores (X-axis) and proficiency scores (Y-axis).

Table 2. Seven highest/lowest similarity scores and their proficiency scores

	Word2Vec						BERT					
	Topic 1			Topic 2			Topic 1			Topic 2		
	PPT	SIM	PRF	PPT	SIM	PRF	PPT	SIM	PRF	PPT	SIM	PRF
Highest	13	0.848	156	32	0.805	181	19	1.000	171	13	1.000	156
	11	0.812	102	12	0.786	154	2	0.997	122	19	0.997	171
	34	0.807	119	13	0.774	156	35	0.993	144	25	0.989	165
	33	0.807	156	11	0.773	102	11	0.993	102	35	0.985	144
	7	0.803	167	19	0.761	171	26	0.991	176	9	0.983	132
	32	0.796	181	18	0.733	172	18	0.979	172	12	0.953	154
	10	0.785	138	21	0.731	186	20	0.977	125	16	0.593	99
Lowest	23	0.625	128	36	0.605	92	34	0.085	119	15	0.096	91
	35	0.616	144	6	0.575	92	28	0.063	148	11	0.073	102
	5	0.597	138	4	0.573	153	14	0.057	167	6	0.069	92
	4	0.595	153	2	0.547	122	9	0.051	132	32	0.064	181
	9	0.590	132	3	0.544	121	4	0.034	153	23	0.029	128
	14	0.543	167	22	0.539	49	27	0.017	120	5	0.016	138
	22	0.482	49	27	0.373	120	7	0.000	167	22	0.000	49

Note. PPT = participant; PRF = proficiency score; SIM = similarity score

References

Crossley, S., Kyle, K., & Dascalu, M. (2019). *Behavior Research Methods*, 51, 14–27.

Dascalu, M., Westera, W., Ruseti, S., Trausan-Matu, S., & Kurvers, H. (2017). In *Artificial Intelligence in Education 2017 Lecture Notes in Computer Science, 10331* (pp. 52–63).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *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)* (pp. 4171–4186).

Lee-Ellis, S. (2009). *Language Testing, 26*(2), 245–274.

Meurers, D. (2015). In *The Cambridge handbook of learner corpus research* (pp. 537–566).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013).