Transformer-architecture-based text similarity and L2 proficiency

One major 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 transformer-architecture neural-network models, which are cutting-edge techniques in machine learning.

Method (Table). Thirty-three L1-Czech L2-Korean learners (age: mean = 24.0; *SD* = 2.69) were asked to write argumentative essays on two topics. Learner proficiency was measured separately using the Korean C-test (Lee-Ellis, 2009; ranging from 0 to 188; mean = 103.74, *SD* = 25.66). Essays from 25 native Korean speakers were collected as a reference text. After electronically converting all the essays with typos and spelling/spacing errors uncorrected, we computed cosine similarity scores between individual learner writing and the reference text by employing two transformer-architecture neural-network models—BERT (Devlin et al., 2018) and GPT-2 (Radford et al., 2019). The similarity scores (predictor) and proficiency scores (outcome) were then submitted to linear regression. In addition, we recruited 10 raters to holistically evaluate learner essays for content, organisation, and language use (as employed in the TOPIK writing evaluation process).

Results (Figure). Whereas the rating–proficiency tendency was uniformly positive, the similarity–proficiency relationships were idiosyncratic. This indicates that the similarity scores, obtained automatically from the neural-network models, fundamentally differ from the rating scores obtained holistically from human evaluation. In addition, no uniform tendency was found across the topics for both models; only a few regression analyses yielded significance for the similarity–proficiency relationships.

This eccentric performance in predicting proficiency indicates two possibilities: (i) the operation of these models may have been greatly influenced by such factors as essay topics in an asymmetric manner; (ii) the transformer-architecture models may not have been adept at extracting a centralised tendency from learner writing in general. Both possibilities are attributable to the properties of the transformer architecture's internal algorithms: they utilise raw sentences (with no POS information) as a basic data-processing unit, assuming that sequences of portions of the sentences comprises a context allowing those sequences to share certain distributions/meanings. Because of this nature, the transformer architecture may not work well in managing learner writing (due to learner language characteristics; Meurers & Dickenson, 2017) compared to its state-of-the-art performance in many downstream NLP tasks.

Together, our findings suggest that the application of NLP techniques to learner corpora needs a researcher's sound understanding of how their algorithms operate in conjunction with various factors that possibly affect their operation.

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

Topic	L2 learner			Native speaker		
	Mean (SD)	Minimum	Maximum	Mean (SD)	Minimum	Maximum
1	95.26 (33.18)	38	158	173.36 (57.85)	91	306
2	91.76 (32.42)	29	156	160.28 (52.41)	81	265

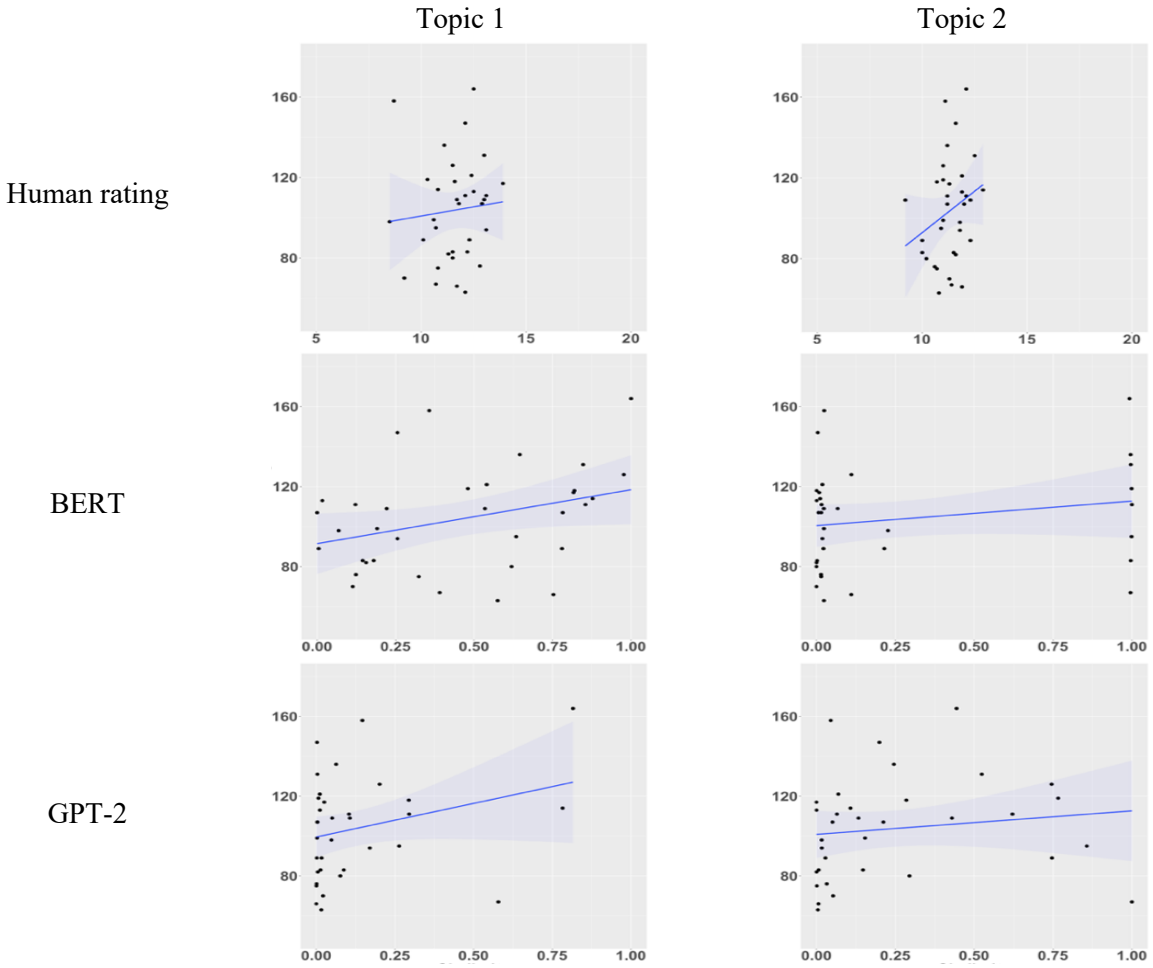

Figure. Rating score (X-axis for human rating) or similarity score (X-axis for BERT & GPT-2) and proficiency score (Y-axis) by model and topic. Statistical significance: $F(1, 32) = 4.017$, $p = .054$, $R^2 = .112$, $B = 26.963$ (BERT, Topic 1).

References

Crossley et al. (2019). *Behavior Research Methods*, 51, 14–27.

Dascalu et al. (2017). In E. *Artificial Intelligence in Education 2017 Lecture Notes in Computer Science, 10331* (pp. 52–63).

Devlin et al. (2018). In *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).

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

Meurers & Dickinson (2017). *Language Learning*, *67*(S1), 66–95.

Radford et al. (2019). *OpenAI Blog*, *1*(8), 9.