NLP-based measurement of text quality for learner writing: Relationship between text similarity and proficiency

Boo Kyung Jung (University of Pittsburgh)

Gyu-Ho Shin (Palacky University Olomouc)

Seongmin Mun (Université Paris Nanterre)

Text similarity

- Two major areas of automatic analysis of learner corpora
  - Lexico-grammatical features (e.g., Kyle, 2016; Lu, 2010)
    - Extracting morpho-syntactic aspects of learner production
    - Substantiating measures/indices that explain/predict characteristics of learner production
  - Text quality (e.g., Burstein et al., 2013; Crossley & McNamara, 2013; Crossley et al., 2019)
    - Semantic-pragmatic aspects of language use in learner corpora
    - Positive relationships between the quality of writing and human raters' evaluation or between proficiency and similarity of spoken production to a test prompt
    - 'Invisible' in nature and thus operationalised
      - Coherence through cohesion devices (e.g., Crossley et al., 2016)
      - The degree of **similarity** relative to the native norm (e.g., Crossley et al., 2019; Dascalu et al., 2017)

Text similarity

- Two major caveats in the area of text quality research
    - Individual roles of these techniques played in text quality measurement need to be clarified with respect to specific constructs of L2 learners' competence
        - Core question: *how does each technique address learner characteristics such as proficiency?*
    - Investigation of applying NLP techniques to text quality assessment of learner corpora occurs in a restricted range of languages, mostly in L2 English
        - Core question: *whether and to what degree do the implications of existing literature hold for non-L2-English contexts, particularly for languages typologically different from English?*

Our study: Aim

- We (i) choose NLP techniques representative of topic modelling and word embedding, (ii) apply each technique to learner writing, and (iii) see if and how similarity scores of learner writing (relative to native speaker writing) explain proficiency
  - Topic modelling
    - Latent Semantic Analysis (LSA)
    - Latent Dirichlet Allocation (LDA)
  - Word embedding
    - Word2Vec
    - BERT

Our study: Methods

- Participants
  - Experimental group: 36 Mandarin-speaking learners of Korean attending one university in Korea (mean age = 24.2; SD = 3.11)
    - Proficiency measured separately through the Korean C-Test (Lee-Ellis, 2009)
  - Reference group: 10 native speakers of Korean (mean age = 27.5; *SD* = 2.93)

- Data collection
  - Two argumentative essays on a separate sheet of paper for 20 minutes each
  - Topics adapted from TOPIK
    - Topic 1: *Which do you think is the most important, preservation vs. exploitation of the nature?*
    - Topic 2: *What affects success the most, competition or cooperation?*

Our study: Methods

- Data processing
  - All the essays were electronically converted into machine-readable format (.txt) files per participant and per topic, with errors and typos uncorrected

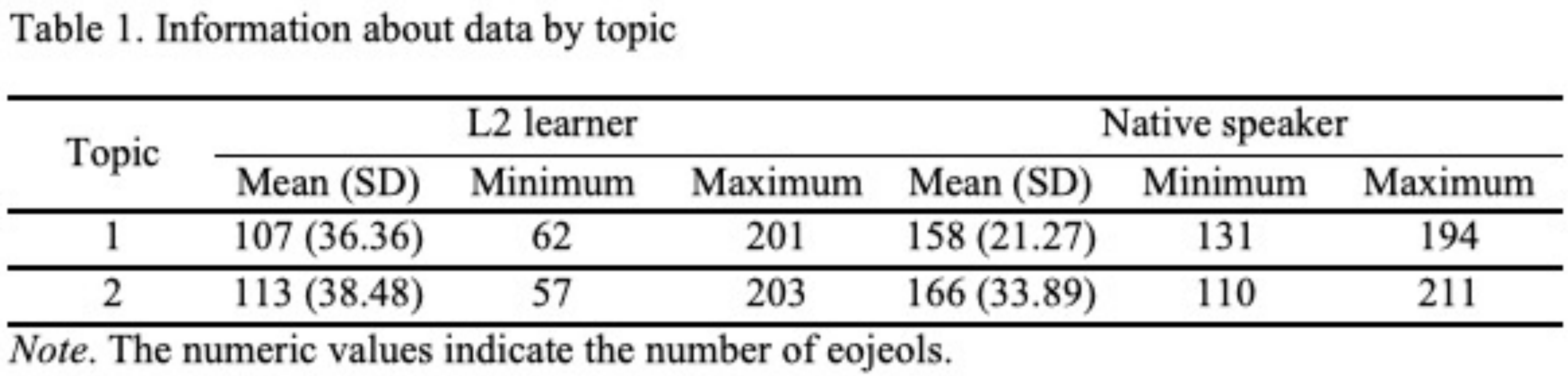

Table 1. Information about data by topic

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

Note. The numeric values indicate the number of eojeols.

Our study: Methods

- Data processing
  - All the function words from the essays were removed and the content words were normalised manually

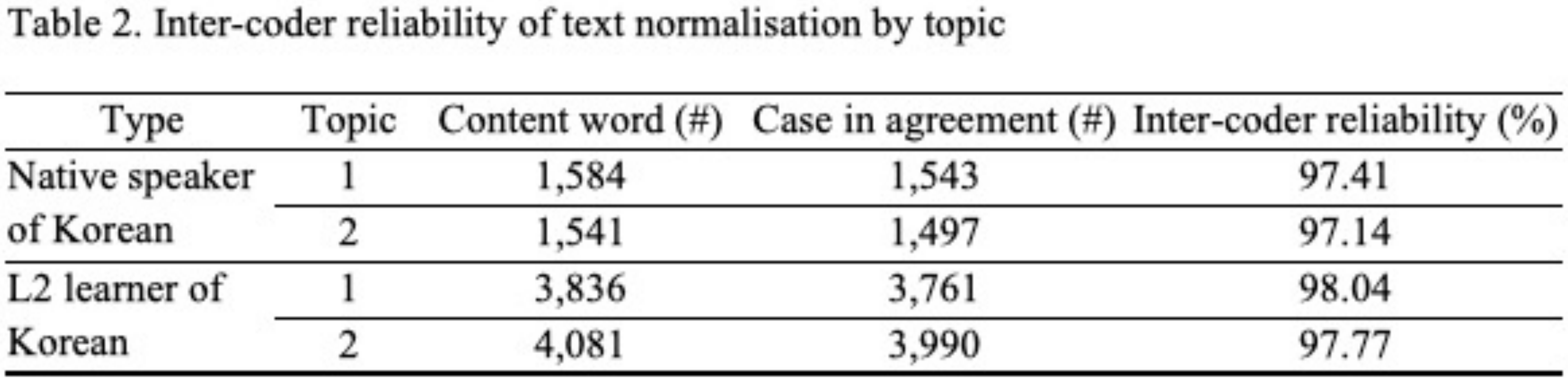

Table 2. Inter-coder reliability of text normalisation by topic

Type	Topic	Content word (#)	Case in agreement (#)	Inter-coder reliability (%)
Native speaker of Korean	1	1,584	1,543	97.41
	2	1,541	1,497	97.14
L2 learner of Korean	1	3,836	3,761	98.04
	2	4,081	3,990	97.77

Our study: Methods

- Data processing
  - Similarity scores between individual learner writing and native speakers’ writing (as a whole) through the cosine similarity

Our study: Methods

- Data processing: LSA
  - Each document was transformed into line-by-line strings per document & converted into a *data.frame* format, comprising rows (representing each document) and columns (representing each word in a document)
  - The pre-processed data were converted into a term-document matrix through *TfidfVectorizer* & dimension reduced by applying *TruncatedSVD* to the matrix (dimension # = 10)
  - 10 topics were extracted from the 10 reduced dimensions per each document, with each of the topics engaged in different magnitude of weight

Our study: Methods

- Data processing: LDA
  - Each document was transformed into line-by-line strings per document & converted into a *data.frame* format, comprising rows (representing each document) and columns (representing each word in a document)
  - A dictionary was generated with unigram words from the whole dataset to determine the data size for model training (topic # = 10)
  - A new data frame was produced with columns of the 10 weight values of each topic and with rows of document numbers

Text similarity: Methods

- Data processing: W2V
  - Each document was transformed into line-by-line strings per document & converted into a *data.frame* format, comprising rows (representing each document) and columns (representing each word in a document)
  - A dictionary was created including all the words in the data & a pre-trained model employed to make the neural network algorithm for the Word2Vec model properly
  - A similarity index model was created to compute cosine similarity between word embeddings given the pre-trained model
  - The index model was inputted to a similarity matrix model to calculate cosine similarity between actual words in the documents

Text similarity: Methods

- Data processing: BERT
  - A data frame was created with rows including individual sentences by essay topics and with columns including each document
    - Label 0: native speaker writing as a whole; Labels 1 to 36: individual learner writing
  - Every sentence in the rows contained [CLS] and [SEP] before and after one sentence, respectively, to indicate sentence boundaries
  - Information extraction process for model training
    - Data labels and tokenised the sentences were extracted to serve as designated indices of the tokens in the pre-trained model
    - Tokens in each sentence were converted into 0 (not attested) or 1 (attested)
    - Information obtained by this process was transformed as tensors

Text similarity: Methods

- Data processing: BERT
  - Model training with GPUs/TPUs in Google Colab & KoBERT as a pre-trained model
    - *eps* = .00000008; *lr* = .00002; *seed* = 42; *batch* = 32; *epoch* = 10
  - Two outcomes
    - A set of 482 arrays (the number of sentences in the dataset)
    - A total of 482 sets of 37 arrays (the number of documents in the dataset)
  - Outliers were excluded from a cluster of values in each document produced by the model
  - Trimmed data were converted into a 2 (document name and centre value per document) by 37 (individual document) matrix to calculate similarity scores

Text similarity: Results

- By-technique results
  - LSA
    - Linear regression
      - Topic 1: *ns*
      - Topic 2: $F(1, 34) = 7.41$, $p = .010$, $R^2 = .179$, $B = 76.87$
    - Despite the graphical tendency, the degree that the similarity scores (by the LSA model) predict the proficiency score varied by topic

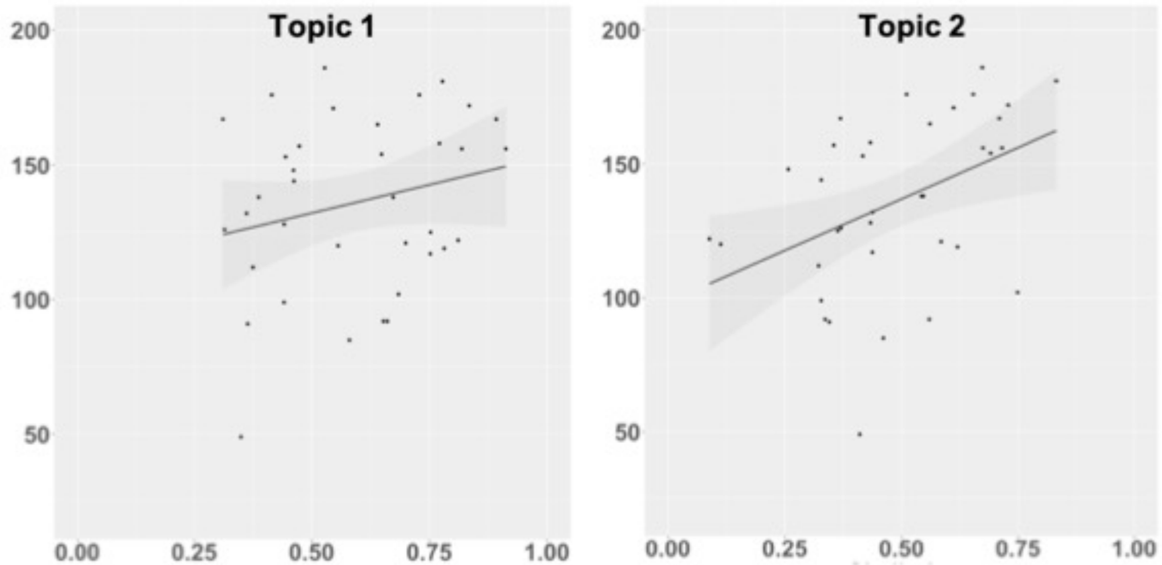

Figure 1. Scatterplot (LSA): similarity scores (X-axis) and proficiency scores (Y-axis).

Text similarity: Results

- By-technique results
  - LDA
    - Most of the similarity scores were bipolarised in their location, regardless of the topics, and many of the scores were around the value of 0
    - Linear regression: *ns*
    - The LDA-based similarity scores did not predict the proficiency scores for the two topics

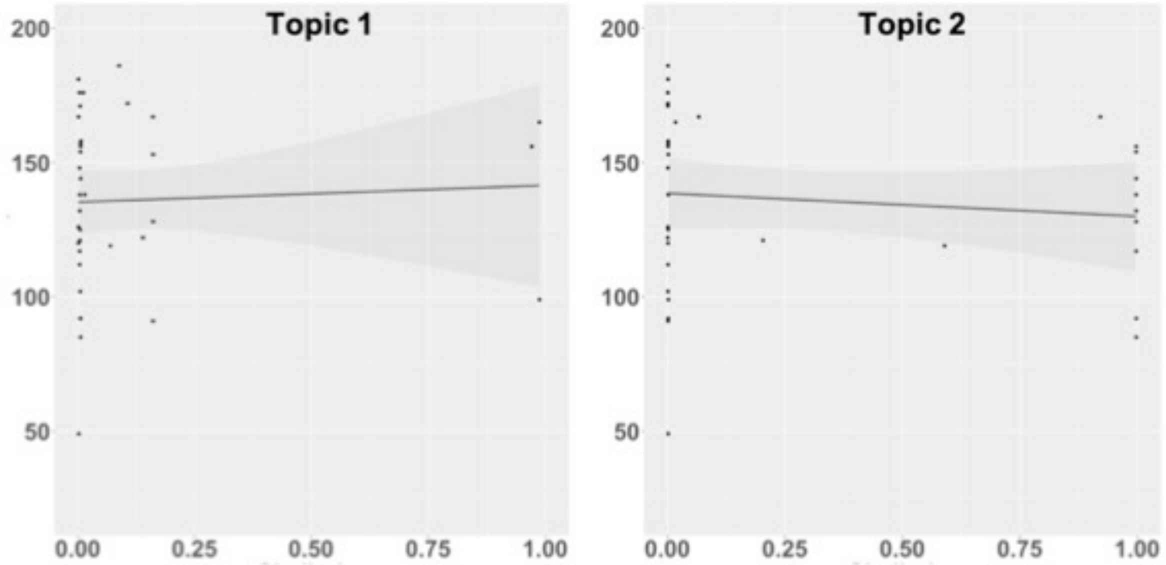

Figure 2. Scatterplot (LDA): Similarity scores (X-axis) and proficiency scores (Y-axis).

Text similarity: Results

- By-technique results
  - W2V
    - Linear regression
      - Topic 1: $F(1, 34) = 3.405$, $p = .074$, $R^2 = .064$, $B = 113.86$
      - Topic 2: $F(1, 34) = 8.748$, $p = .006$, $R^2 = .181$, $B = 172.59$
    - The Word2Vec-based similarity scores well-predicted the proficiency scores, the degree to which being influenced by the topics

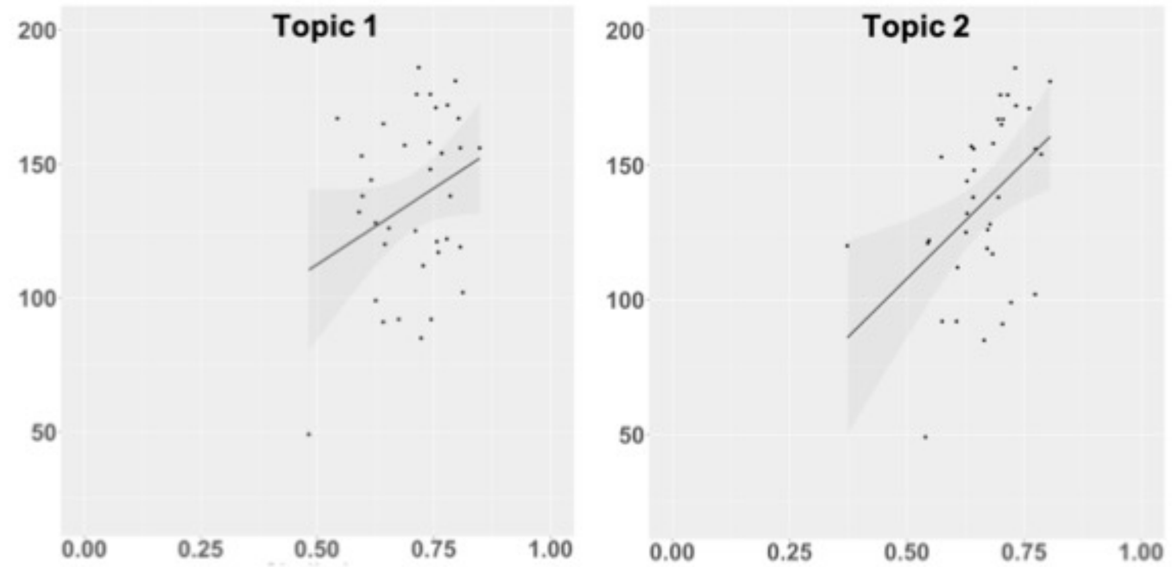

Figure 3. Scatterplot (Word2Vec): Similarity scores (X-axis) and proficiency scores (Y-axis).

Text similarity: Results

- By-technique results
  - BERT
    - Linear regression
      - Topic 1: *ns*
      - Topic 2: $F(1, 34) = 3.79$, $p = .060$, $R^2 = .074$, $B = 32.296$
    - Topic sensitivity when it comes to the model performance of BERT

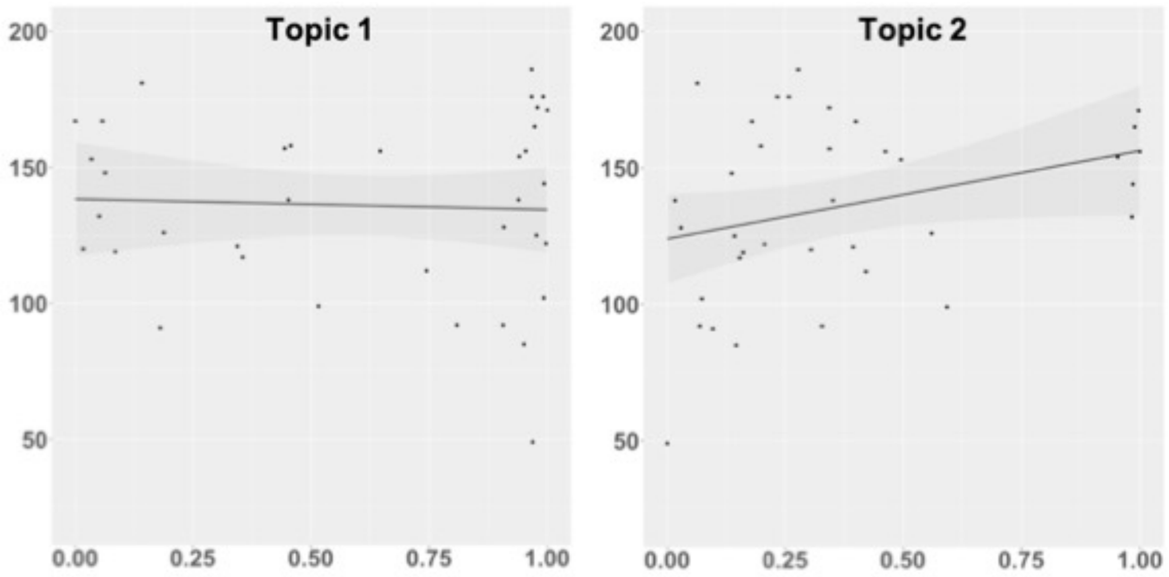

Figure 4. Scatterplot (BERT): Similarity scores (X-axis) and proficiency scores (Y-axis).

Text similarity: Results

- Between-technique results

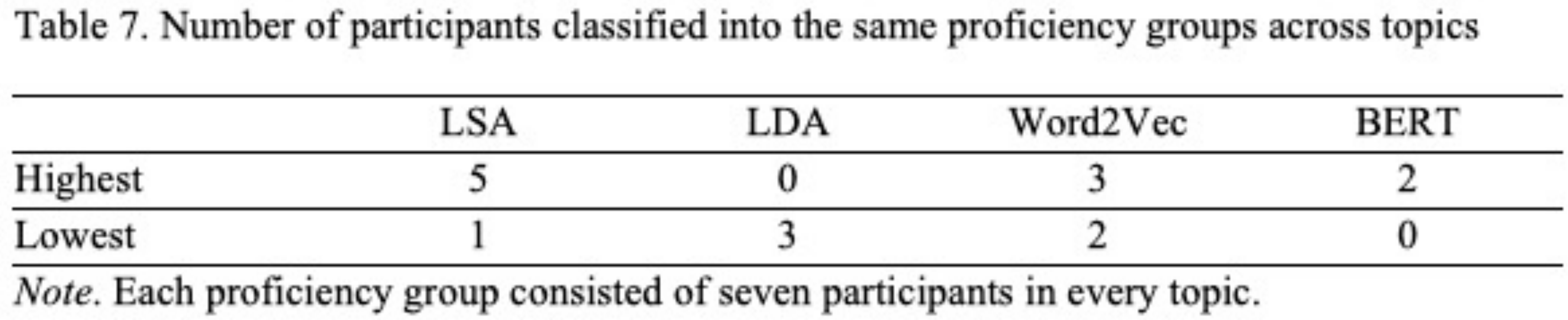

Table 7. Number of participants classified into the same proficiency groups across topics

	LSA	LDA	Word2Vec	BERT
Highest	5	0	3	2
Lowest	1	3	2	0

Note. Each proficiency group consisted of seven participants in every topic.

  - Although caution must be taken in interpreting these results due to a small number of participants involving each proficiency group, these results indicate that not all NLP techniques for text similarity are equally good for explaining proficiency

Text similarity: Results

- Between-technique results

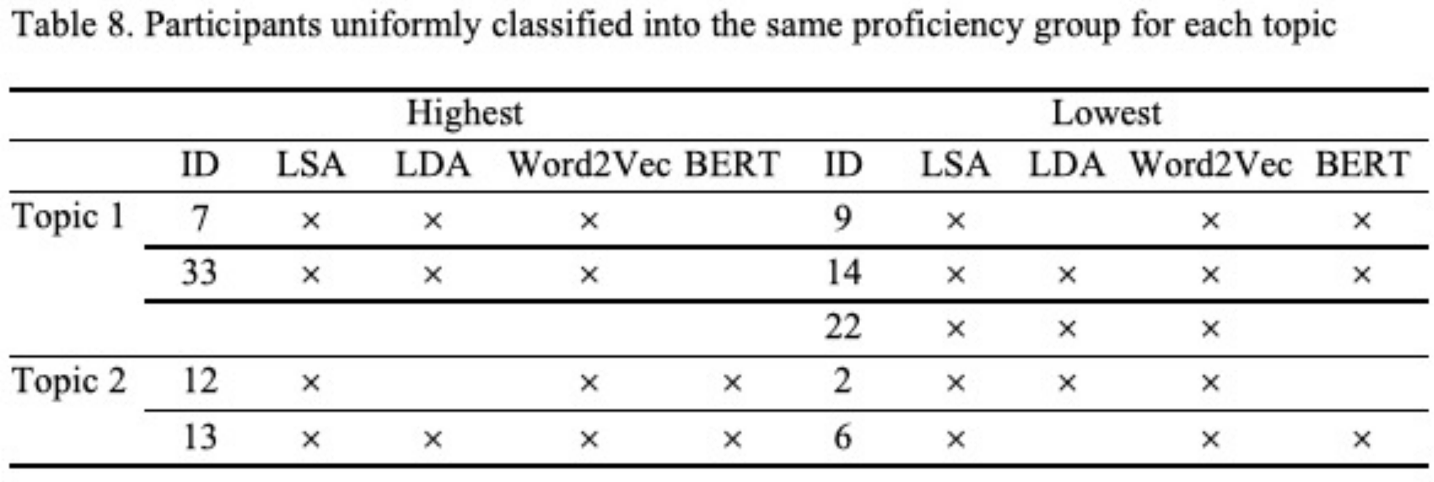

Table 8. Participants uniformly classified into the same proficiency group for each topic

	Highest					Lowest				
	ID	LSA	LDA	Word2Vec	BERT	ID	LSA	LDA	Word2Vec	BERT
Topic 1	7	×	×	×		9	×		×	×
	33	×	×	×		14	×	×	×	×
						22	×	×	×	
Topic 2	12	×		×	×	2	×	×	×	
	13	×	×	×	×	6	×		×	×

  - Overall, the highest group used content words which were also frequently attested in the native speakers' essays more often than the lowest group (+ more spelling errors)

Text similarity: Implications

- The results show…
  - Asymmetric degrees to which the similarity scores of each technique explained the proficiency scores
    - W2V > LSA and BERT > LDA
  - Model performance sensitive to essay topics (and particularly to word use)
  - Global limitation to capturing individual variations involving learner writing

Text similarity: Implications

- What do the results suggest?
  - Given the specifications involving each technique, the applications of these techniques to learner corpora need to be based on how the algorithms of these NLP techniques operate in conjunction with the characteristics of learner language

  - One general limitation to the current NLP techniques for text similarity (and beyond): they still rely on words, possibly falling short of incorporating contexts (as a genuinely linguistic sense) identified through semantic-pragmatic features in the course of data processing

References (selected)

Burstein, J., Tetreault, J., Chodorow, M., Blanchard, D., & Andreyev, S. (2013). Automated evaluation of discourse coherence quality in essay writing. In M. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 267–280). New York: Routledge.

Crossley, S., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning and Technology, 17*(2), 171–192.

Crossley, S., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, 51, 14–27.

Crossley, S., Kyle, K., & McNamara, D. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods, 48*(4), 1227–1237.

Dascalu, M., Westera, W., Ruseti, S., Trausan-Matu, S., & Kurvers, H. (2017). ReaderBench learns building a comprehensive automated essay scoring system for Dutch language. In E. André, R. Baker, X. Hu, M., Rodrigo & B. du Boulay (Eds.). *Artificial Intelligence in Education 2017 Lecture Notes in Computer Science, 10331* (pp. 52–63).

Kyle, K. (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.

Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch Analysis. *Language Testing, 26*(2), 245–274.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*, 474–496.

Thank you for your listening!