Issues in the Evaluation and the Measurement of Learner Language

Kondo, Yusuke

(Citation)
Learner Corpus Studies in Asia and the World,5:95-104

(Issue Date)
2020-12-21

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81012492

(URL)
https://hdl.handle.net/20.500.14094/81012492

Learner Corpus Studies in Asia and the World, vol. 5 (Papers from LCSAW2020) ISSN: 2435-2632
Published by School of Language & Communication, Kobe University, Japan

Issues in the Evaluation and the Measurement of Learner Language

Yusuke KONDO
Waseda University

Abstract

To construct learner corpus is a project that brings together different fields of study and knowledge. Although second language acquisition, language testing, automated scoring, and task-based language teaching/learning benefit from learner corpus research, learner corpus research does not fully refer to the research results in its adjacent fields. The purpose of this study is to identify common issues in automatic scoring of second language (L2) and learner corpus research and to point out the need for knowledge from their adjacent disciplines to build a more useful learner corpus. The issues presented are found in the evaluation and the measurement of learner language. After reviewing the basic concept of automated scoring, issues of predictor variables and human rating are presented in automated scoring, and then, issues of features and data collection time period are mentioned in a longitudinal learner corpus research. One of the ways to solve the issues presented in this paper is the collaboration among researchers in automated scoring and in learner corpus, expertise in language testing and in task design. This collaboration gives useful insights to the researchers and the expertise in these fields

Keywords

Automated scoring, Learner corpus research, Human rating, Linguistic features

1. Introduction

To construct learner corpus is a project that brings together different fields of study and knowledge. Second language acquisition, language testing, automated scoring, and task-based language teaching/learning, these disciplines benefit from learner corpus research. However, each of the disciplines does not fully refer to the research results in its adjacent fields. The purpose of this study is to identify common issues in automatic scoring of second language (L2) and learner corpus research and to point out the need for

knowledge from their adjacent disciplines to build a more useful learner corpus. The issues presented are found in the evaluation and the measurement of learner language. Firstly, the basic mechanics of automatic scoring is reviewed, and then issues are identified that need to be solved in automatic scoring research. Secondly, a study of longitudinal corpus is presented to show that the similar issues as in automated scoring research can be found in learner corpus research.

2. Basic concepts of automated scoring of L2

Suppose a situation that a teacher scores essays of two students'. Student A are given 4 out of 5, and his essay consists of 520 words and includes 7 grammatical errors. Students B are given 2 out of 5, and her essay consists of 395 words and includes 12 grammatical errors. In this case, the scores can be expressed by using the numbers of words and errors.

Student A's case

$$520 \times x + 7 \times y = 4$$

Student B's case

$$395 \times x + 12 \times y = 2$$

In the formulae above, x is the index of how the teacher think the number of the words to be important in essays; and y, the number of grammatical errors. What score will the teacher give to Student C's essay that consists of 455 words and includes 4 grammatical errors? By using the two formulae for Students A and B, we can obtain x and y. Then, we can predict the score for Student C's essay by assigning these values, x and y, to the formula below.

Student C's case

$$455 \times x + 4 \times y$$

This is the basic concept of automated scoring of L2 writing. We obtain the importance of the number of words and grammatical errors, x and y in the formulae for the essays of Students A and B, and predict the score of Student C's essay.

However, to construct an automated scoring system is not such an easy task. We sometimes encounter cases below.

Students D's case

$$450 \times x + 6 \times y = 4$$

Student E's case

$$450 \times x + 6 \times y = 5$$

In these cases, both students wrote 450 words and make 6 grammatical errors, but they received different scores. This can be caused by features chosen to predict scores and/or unreliable rating.

3. Unreliable rating

If we are given a score by an unreliable rater, the score is unreliable. This section shows unreliable rating detected in the process of constructing an automated scoring system in Kondo and Ishii (2017). In this study, the learners completed oral discourse completion tasks. Below is an example of the tasks.

When you want to end your conversation, what would you say in the conversation?

A total of 14 English language teachers gave scores 1 to 4 to about 500 learners speech. Score 1 indicates no utterance; Score 2, "inappropriate utterance;" Score 3, "inappropriate utterance but can understand the speaker's intention;" and Score 4, "appropriate utterance." The 14 teachers, who joined as raters in this study, received a short rater training where they were given the evaluation criteria and sample speeches with scores. The number of speeches that an individual rater scored are different. Some raters scores 100 speeches; and other raters, 30. Overlapping speeches at the rate of 20% were included in each of the rating in order to examine the internal consistency of the ratings. For example, if a rater scores 5 speeches, they are A, B, C, D and A. The rater scores Speech A twice. The internal consistency, the degree of agreement was calculated in these overlapping speeches. Table 1 shows the number of subjects whom a rater gave scores, Cohen's kappa, and the exact agreement ratios. Perfect agreements are found in the ratings by Raters 1 and 11, but, on the other hand, poor agreements are found in the ratings by Raters 6, 7, and 14. Even if the performances are the same, different scores can be given to them. This is one of the problems to be solved to construct automated scoring systems for L2. In the model of automated scoring, the score is a criterion variable that is reliable.

Table 1

The number of subjects, Cohen's kappa, and the exact agreement ratios

	N of subjects	κ	% of exact agreement
Rater 1	50	1	1
Rater 2	67	0.89	0.98
Rater 3	100	0.71	0.83
Rater 4	100	0.92	0.96
Rater 5	150	0.96	0.98
Rater 6	50	0.56	0.76
Rater 7	17	0.32	0.64
Rater 8	100	0.95	0.99
Rater 9	100	0.56	0.8
Rater 10	150	0.81	0.9
Rater 11	117	1	1
Rater 12	150	0.77	0.91
Rater 13	150	0.89	0.94
Rater 14	100	0.62	0.75

4. Wrong features

Suppose that we construct an automated scoring system that predicts scores in an essay by using two linguistics features: the number of words and grammatical errors. However, if the instructions in the target essay includes "No limit to the number of words" and "the number of errors does not deduct your score," then, the prediction accuracy of the automated scoring system must be very low. In the examination of linguistic features as a predictor variable, it is important to understand the instructions of the task that we introduce automated scoring system to.

In Cucchiarini, Strik, and Boves. (2000), a study on automated scoring, phoneticians evaluated the read-aloud speeches of L2 learners of Dutch in terms of overall pronunciation (OP), segmental quality (SQ), fluency (FL), and speech rate (SR) with 10-point scale. Table 2 shows the correlation coefficients between these four sorts of scores and three linguistic features: time duration (TL), rate of speech (ROS), and likelihood ratio (LR). LR is an index of similarity between learners' speech and model speech

(Native speakers of Dutch) calculated by automated speech recognition system. The majority of the correlation coefficients in Table 2 are plausible results. For example, the correlation coefficients between FL and SR and ROS are fairly high. However, we can find moderate correlation coefficients between OP and LR and between SQ and LR. Because LR is the index of similarity with the native speakers' pronunciation, these correlation coefficients must be higher than the actual ones as in the correlations between SR and ROS. In this case, we can decide neither that the raters are unreliable nor that LR is a wrong feature to predict the scores of OP and SQ.

Table 2

Correlation coefficients between scores and linguistic features in Cucchiarini, Strik, and Boves (2000)

	OP	SQ	FL	SR
TD	-0.79	-0.75	-0.91	-0.90
ROS	0.82	0.79	0.93	0.92
LR	0.49	0.45	0.55	0.59

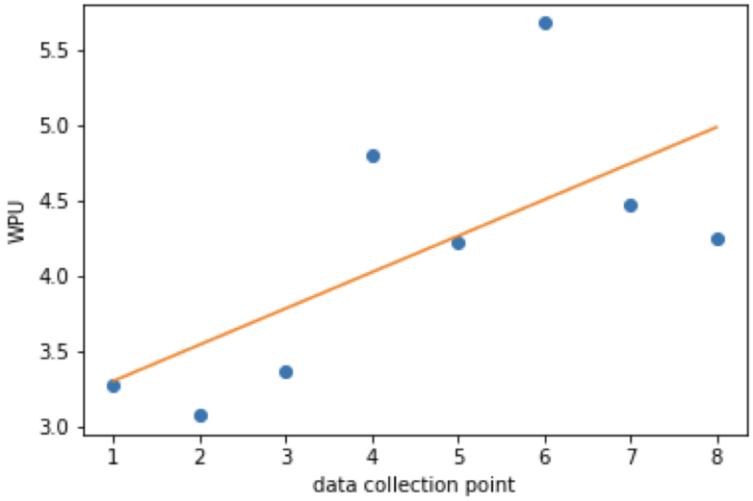
5. Sharing issues in common with Learner Corpus Research

As mentioned above, to construct a model of automated scoring system is to find the relationship between features and scores and apply it to predict a score of writing of speech. In other words, to construct the model is to elicit a formula from the existing data and to apply it to predict scores. The issues to be solved is to establish reliable ratings and find accurate predictors in features. This sort of issue is found in studies in learner corpus research.

In a longitudinal corpus study, a growth curve model is employed to capture the development of learner language. The idea is to consider a regression line with the data collection point as x and the feature as y as the true developmental curve. Figure 1 shows fictitious data of an individual development in speech rate. The learner takes the same speaking test every year from the first year of junior high school to the second year of college, with y being the feature on speech rate and x being the times of the tests. As in the case of the construction of automated scoring system, we examine features and data collection points: whether the features can capture the development, whether the intervals between the data collection points are appropriate, and whether the data collection is conducted in an adequate period for capturing the development.

Figure 1

Fictitious data of individual development in speech rate



Abe, Kondo, Fujiwara, and Kobayashi (2020) investigated the development of complexity in learner language. The study analyzed learner languages in a longitudinal learner corpus. A total of 104 Japanese learners of English, high school students, took a monologue speaking test called Telephone Standard Speaking Test (ALC Press, 2016) eight times during three years from 2016 to 2018. Although this study examined several features in complexity, this section focuses on the development of the syntactic complexity by utilizing a feature, degree centrality.

Degree centrality is an index to be often used to calculate the complexity of networks. All the nodes in Figure 2 are directly connected. This can be said to be a flat network. In Figure 3, on the other hand, the nodes are only connected to their neighboring nodes. This is called a deep network. The degree centrality of the network in Figure 2 is 1; and that in Figure 3, 0. In this study, the index, degree centrality is applied to measure syntactic complexity. All the utterances in this study were transcribed and parsed on the basis of dependency grammar by spaCy (Honnibal and Montani, 2020).

When the phrase "in my pajamas" is an adjectival phrase in the sentence, "I shot an elephant in my pajamas," the sentence is parsed as in the left tree in Figure 4, but, on the other hand, "in my pajamas" is an adverbial phrase, it parsed as in the right tree in Figure 4. The left tree has a deeper structure than that in the right one. In this study, this deepness is regarded as the syntactic complexity, and it is expected that the complexity should develop with each speaking test.

Figure 5 shows growth curves of the syntactic complexity of the learner language in this longitudinal corpus. To demonstrate growth, the index is standardized and the values of 1 minus the degree centrality are plotted in this graph. The bold line in the graph is the line with the average slope and intercept. According to the graph, the degree centrality varies greatly between learners, and almost no change is found in the index (the average of the slopes is 0.1). There are two ways to interpret the results. Firstly, it was not appropriate time period to capture the development of the syntactic complexity. In other words, the syntactic complexity cannot change from the first to the third grade in high school. Secondly, Degree centrality is not an appropriate feature to measure the syntactic complexity. We cannot choose the one. We need to find out appropriate time period and an appropriate feature to recognize the syntactic complexity.

Figure 4

Difference in Deepness Between Two Interpretations of a Sentence

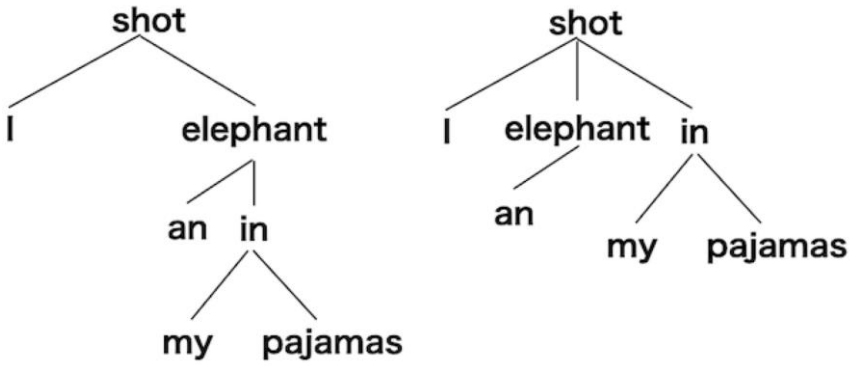
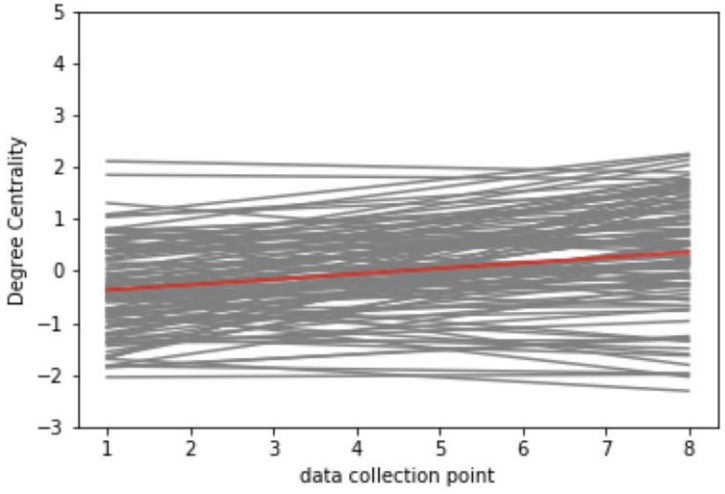


Figure 5

Growth Curves of the Syntactic Complexity



6. Task design

Another problem to be solved both in studies of automated scoring and learner corpus research is task design. In view of language testing, if raters have difficulty to decide to give a performance to 4 or 5 in 5-point scale, which is considered to be attributed to task design, then it is not a good task. If human raters have difficulty to score performance, it is also difficult for automated scoring system. In a longitudinal learner corpus, if researchers find no difference in performance between data collection points 1 and 2, the researchers can assume the task to be one of the reasons why they do not capture the difference.

Think about the task where Students A and B are given two similar pictures individually, ask questions to find the differences between the two pictures. The pictures describe a family getting together happily in their living room. A student may ask a question such as "What is on the table?" and "Is the Christmas tree beside the fireplace?" It is a good task to observe the variety of the correct use of prepositions. Raters can discriminate good preposition users with poor users. Leaner corpus researchers can find the development of the correct use of prepositions. However, it may be fairly difficult to find out the development of the syntactic complexity in this task if students do this task several times. The syntactic complexity cannot be evaluated in this task.

7 Conclusion

In this paper, common issues in automated scoring and learner corpus research are identified: unreliable ratings, wrong features, and inappropriate time period. Furthermore, the importance of task design is pointed out in order to collect the data of learner language. One of the ways to solve the issues presented in this paper is the collaboration among researchers in automated scoring and in learner corpus, experts in language testing and in task design. This collaboration gives useful insights to the researchers and the experts.

References

Abe, Kondo, Fujiwara, and Kobayashi (2020). A longitudinal study of novice learners' development of complexity. Teaching and Learner Corpora 2020. Online Conference. https://langident.hypotheses.org/talc2020/talcmainconference

ALC Press (2016). Telephone Standard Speaking Test (TSST). Retrieved from https://tsst.alc.co.jp/biz/en/

Cucchiarini, C., Strik, H., & Boves, L. (2000). Different aspects of expert pronunciation

quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication*, *30*, 109-119.

Honnibal, M., & Montani, I. (2020). spaCy (Version 2.3.2) [Computer software]. https://spacy.io/

Kondo and Ishii (2017). Presenting practicality of automated speech scoring system in English language education program. *Language Education & Technology*, *54*. 23-40.