Measuring Similarity to Observe Learners' Syntactic Awareness in Web-Based Writing Environments

Kang, Min

Kawamura, Koichi

Shao, Shuai

Kashiwagi, Harumi

Ohtsuki, Kazuhiro

Measuring Similarity to Observe Learners' Syntactic Awareness in Web-Based Writing Environments

Min Kang[1][0000-0003-2211-153X], Koichi Kawamura[1], Shuai Shao[1], Harumi Kashiwagi[2], and Kazuhiro Ohtsuki[1]

[1] Graduate School of Intercultural Studies, Kobe University. Kobe 657-8501, Japan
[2] Institute for Promotion of Higher Education, Kobe University. Kobe 657-8501, Japan
kang@kobe-u.ac.jp

Abstract. Writing in a foreign language is a struggle for learners and revising their writings is time consuming for teachers as well. For this reason, writing support systems have been widely proposed and one of its main functions is to automatically detect and revise errors in learners' writings. However, the detection technologies are a work in progress and the effectiveness of error revision feedback is arguable. Meanwhile, numerous efforts have been made to enhance learners' writing proficiency and reduce errors. Reading is considered as one of the important strategies. However, few studies have reported the linguistic knowledge that learners pay attention to and how they use the knowledge of web-based learning in their writings. In this paper, we performed a reading-to-write experiment in a web-based writing environment and analyzed reading materials and learners' writings to explore how to observe learners' awareness of syntactic structures in materials. Sentence patterns, proposed in our previous studies, have been introduced to categorize sentences, and the syntactic similarities between reading materials and learners' writings have been calculated. The experimental results revealed that students showed higher comprehension of content but displayed poor attention towards syntactic structures in reading activities, if the structures were not significantly salient. It is assumed that the similarity measure is effective in observing students' awareness of syntactic structures in materials, and further studies are needed to automatically observe the awareness.

Keywords: Syntactic Awareness, Reading-to-write, Similarity Measure, Web-Based EFL Writing.

1 Introduction

Writing in a foreign language is a struggle for learners. Besides organization and clarity in content, writing necessitates the accurate use of lexical and syntactic knowledge [1]. Hence, grammatical revision is an important process in foreign language writing. However, revising learners' writings is time consuming for teachers. For this reason, automatically detecting and revising errors in learners' writings is a popular research topic not only in the field of educational technology but also in natural language pro-

cessing [2]. Writing support systems have been proposed [3-6]. Nevertheless, detection technologies for this purpose are still a work in progress and the effectiveness of error revision feedback is arguable [7-9].

On the other hand, considerable effort has been put in teaching writing to enhance learners' writing proficiency. This means that errors in learners' writings decrease as proficiency improves. Reading is regarded as one of the important strategies of enhancing writing proficiency. Reading and writing are interdependent and writers' linguistic skills, contextual awareness, and strategies, etc., are influenced by the information in source texts, in writers' prior experiences and learning etc. [10]. It has been stated that learners' writing skills concerning content, organization, vocabulary, and language use are associated with their reading skills [11]. Furthermore, the visual-syntactic text formatting technology that visualizes syntactic structures has been experimentally used on reading to enhance syntactic awareness. The experimental results clarified that the technology raised students' awareness of syntactic structures, and the written conventions and writing strategies of low-proficiency students were significantly influenced by the technology [12]. However, problems related to learners' awareness of linguistic knowledge in source texts and how the knowledge provides a scaffold of support to writing, emerged. We suggest that, in a web-based language learning environment, observing learners' attention towards linguistic knowledge in source texts is necessary as well. Based on the observation, a scaffold of support can be provided to writing at last.

In this paper, we aim to present an approach to automatically and quantitatively observe the correlation between the reading and writing activities. We focus the observation on learners' awareness of syntactic structures in reading materials. Although literature [12] has provided a technical method of raising syntactic awareness, it has not been reported how to automatically and quantitatively observe learners' awareness in a web-based language learning environment.

We consider that learners' awareness can be observed by measuring similarities between reading materials and learners' writings. Therefore, we aim to perform an experiment that includes reading-to-write tasks in a web-based writing environment and analyze reading materials and learners' writings. Sentence patterns, proposed in our previous studies, are introduced to categorize sentences, and the syntactic similarities between reading materials and writings are calculated [13]. In the next section, we explain the details of the experiment. We propose the method of similarity measure in Section 3, and then provide the experimental results and discussion in Section 4.

2 Experimenting with Reading-to-Write Tasks

2.1 Web-based writing environment

Generally, there are two kinds of reading styles related to writing in classroom learning: reading-to-write and reading-to-integrate. We adopted the reading-to-write style in the experiment to investigate learners' awareness. Many studies have focused on the relationship between reading-to-write and writing. Most claimed that reading-to-

write strongly influences writing [14]. Although it is evident that integrated reading improves learners' writing proficiency, it is a harder task to observe learners' awareness.

We developed a web-based writing system to perform the experiment. Two webpages were designed to provide two reading-to-write tasks. Each page included a paragraph essay on the top followed by two related questions. The first question asks the participant if he/she has read the essay. The second asks to write a response essay in relation to the paragraph essay. An input space for writing is given below the second question. Meanwhile, in order to clarify if learners are sensitive to salient syntactic structures, we colored the present tense verbs in third person singular, which appear on the second page, red [15].

2.2 Reading materials

As the study focuses on how to observe learners' awareness, easy-to-read materials were used to reduce comprehension difficulties and errors in writing. Two paragraphs were chosen from a text book for the freshmen of Kobe University. The essays consist of 156 words (13 sentences) and 152 words (15 sentences) with the topics focusing on bosses in offices and future jobs, respectively. Hence, the questions related to the response essays on the first page were as follows:

- Question 1: Have you read the paragraph before?
- Question 2: Please write a short essay on your boss.

2.3 Participants and procedure

There are 12 participants consisting of second-year, third-year, and senior students of Kobe University, with a major in global culture. This makes the reading materials easy to read for them.

The students were required to log in to the web-based writing system, and then complete the two reading-to-write tasks without using a dictionary, in an orderly manner. To avoid losing participants' attention, the essays are limited to 5 sentences or 70 words so that the experiment takes around fifteen to twenty minutes.

3 Similarity Measure

Learners' awareness can be observed by measuring similarities between their writings and reading materials based on the hypothesis that a learner tends to imitate the syntactic structures he/she pays attention to during reading.

Numerous researches in the field of natural language processing have addressed the issue of similarity measures for semantic or syntactic analysis. Recently, Gali et al. proposed a framework for syntactic analysis [16]. Although we adopted the tree Levenshtein distance to measure syntactic similarities and failed to interpret the results,

it is important to further investigate the similarity measures for analyzing learners' awareness.

In this paper, we measure syntactic similarities between documents by categorizing sentences according to sentence patterns and then calculating Euclidean distances.

3.1 Sentence patterns

Table 1 gives the sentence patterns that learners are required to know at the start of English language learning [13]. In the column "Pattern description," we describe the features of the patterns. It is a common feature in patterns that a sentence is differentiated only by the subject or the verb of the sentence. In addition, each sentence pattern has four sub-classifications which are combinations of tense and polarity: a) a present tense affirmative sentence pattern (pre_aff), b) a present tense negative sentence pattern (pre_neg), c) a past tense affirmative sentence pattern (past_aff), and d) a past tense negative sentence pattern (past_neg). The pattern names will be used in the next section.

The subject-verb phrase of a sentence that starts with the subject and ends with the verb can be easily extracted by the use of a dependency parser. We used the Stanford Parser to extract subject-verb phrases of sentences [17].

Table 1. Sentence patterns

Pattern names	Pattern description
P1	A subject is the first person "I."
P2	A predicate verb is am/is/are/be/have/has/exist/ exists.
P3	A predicate verb is think/believe/consider/guess/ suppose/assume.
P4	Can/be able to/am able to/is able to/are able to is included in a subject-verb phrase.
P5	A subject-verb phrase is excluded from the above patterns.

3.2 Similarities between two documents

We calculated the Euclidean distance of the two documents as similarities. For each document, all the sentences are categorized within the sentence patterns. The ratios of the patterns used in the document are calculated. Then, the Euclidean distance between the two documents is calculated by using the ratios, and is defined as the similarity. Understandably, the larger the value, the lower the similarity.

4 Results and Discussion

We collected 35 sentences and 34 sentences in the two reading-to-write tasks, respectively. There are 13 sentences in the first reading material (Reading_A) and 15 in the

second reading material (Reading_B). Here, simple sentences, complex sentences, and complicated sentences are included.

First, we summarized students' writings. Then, we categorized all the sentences in each student's writings, Reading_A and Reading_B, and calculated the Euclidean distances of the writings and the materials.

4.1 Summary of students' writings

There are 11 students who answered "No" to Question 1 in the two tasks which indicated that it was the first time the students had read the materials. There were a few syntactic errors and spelling mistakes in their writings but all were easy to read. The writings showed high topic similarity as well. Therefore, the students had sufficient reading proficiency and writing proficiency concerning the reading-to-write tasks.

Conversely, it seems that students did not notice the display difference in the present tense verbs in third person between Reading_A and Reading_B, or they did not pay attention to the use of the present tense verbs in third person. There are 7 verbs ending in *–s* in Reading_A and 7 red verbs ending in *–s* in Reading_B. In the reading-to-write task related to Reading_A, 4 present tense sentences in third person singular were used in students' writings and one error was found. In the reading-to-write task related to Reading_B, there were 11 present tense sentences in third person singular while there were 5 verbs without *–s*. The error percentage in the second task is larger. It is thus believed that coloring does not raise students' awareness.

4.2 Observing students' awareness by measuring similarities

Table 2 shows the Euclidean distance values. The values related to Reading_A vary from 0.44 to 1.19, and the distances corresponding to Reading_B range from 0.21 to 0.90. Here, S1-S12 denote the 12 students, and the values are ordered by those in Reading_A. The syntactic structures in the writings of S1, S6, and S7 are very similar to that of Reading_A in comparison to S9's essay. The structures in S11's essay are most similar to Reading_B and those of S3 are far different.

Table 2. Euclidean distance values

Students	S1	S6	S7	S4	S8	S12
Reading_A	0.44	0.44	0.44	0.60	0.71	0.71
Reading_B	0.76	0.52	0.43	0.70	0.43	0.25
Students	S11	S3	S10	S5	S2	S9
Reading_A	0.74	0.83	0.85	0.88	1.05	1.19
Reading_B	0.21	0.90	0.52	0.52	0.56	0.38

We drew the distributions of sentence patterns used in the essays of S1, S6, and Reading_A in Fig. 1 and the distributions of S9 and Reading_A in Fig. 2. Here, the suffixes "Pre_Aff" and "Past_Aff" mean a present tense affirmative sentence pattern

and a past tense affirmative sentence pattern, respectively, as mentioned in Section 3.1.

It can be observed in Fig. 1 that S1 and S6 tend to reuse sentence patterns appearing in Reading_A. Although the ratios of the patterns used in the students' writings are very different from the material, the sentence patterns are limited to those appearing in Reading_A. Students may be potentially or consciously influenced by syntactic structures in the material on reading.

However, Fig. 2 indicates that S9 did not refer to the patterns. There is no reuse of sentence patterns in Reading_A and the complete essay consists of past tense affirmative sentences with be verbs. Obviously, the student ignored the sentence patterns in Reading_A.

Because the results from the task for Reading_B were similarly inclined, as mentioned above, we omitted the figures corresponding to the second task.

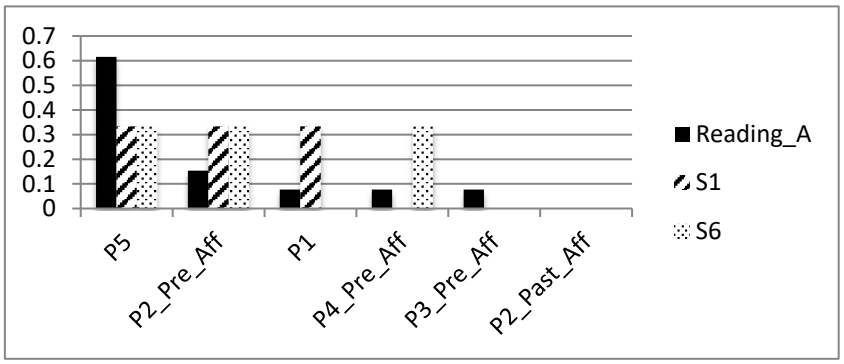

Fig. 1. The distribution of sentence patterns with high similarities

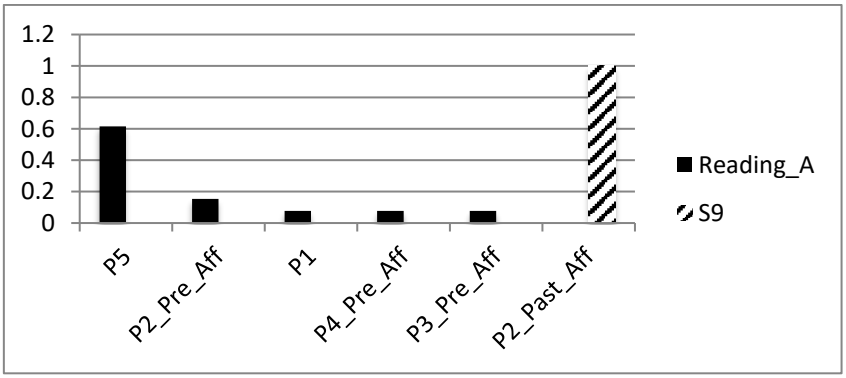

Fig. 2. The distribution of sentence patterns with low similarities

Therefore, it is noticed that a Euclidean distance value based on sentence patterns can prove if learners tend to reuse syntactic structures in reading materials. If the value is small, it means a learner may be aware of syntactic structures on reading. Fur-

thermore, by verifying the distribution of the structures appearing in learner's writings, we may find which structure the learner tends to reuse on writing. The results claim that measuring similarity by categorizing sentences according to sentence patterns may be an effective approach to automatically observe learners' syntactic awareness.

5 Conclusion

In this paper, we performed a reading-to-write experiment in a web-based writing environment and analyzed reading materials and learners' writings. Sentence patterns that were proposed in our previous studies were introduced here to categorize sentences, and the syntactic similarities between reading materials and writings were calculated. The experimental results revealed that most of the students showed higher comprehension on topics but poor attention towards syntactic structures in reading activities, despite parts of the structures being colored. It is assumed that the similarity measure is effective in observing students' awareness of syntactic structures in materials.

On the other hand, the students involved in the experiment are limited and the essays are short. Therefore, we need to improve such points to enhance the precision of the approach. Further studies are still needed to automatically observe learners' awareness.

Acknowledgement

This work is supported by JSPS KAKENHI Grant Number JP17K01081.

References

1. Kimura, H., Kimura, T., Shiki, O.: Theory and practice in reading and writing: nurturing independent learning. TAISHUKAN Publishing Co., Ltd., Tokyo (2010) (in Japanese).
2. Leacock, C., Chodorow, M., Gamon, M., Tetreault, J.: Automated grammatical error detection for language learners. 2nd Ed. Morgan & Claypool Publisher (2014).
3. Bowerman, C.: Writing and the computer: an intelligent tutoring systems solution. Computers & Education 18(1-3), 77-83 (1992).
4. Yeh, S., Lo, J.: Using online annotations to support error correction and corrective feedback. Computers & Education 52(4), 882–892 (2009).
5. Kunichika, H., Koga, T., Deyama, T., Murakami, T., Hirashima, T., Takeuchi, A.: Learning support for English composition with error visualization. The IEICE Transactions on Information and Systems (Japanese edition) 91(2), 210-219 (2008).
6. Wilson, J., Czik, A.: Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. Computers & Education 100, 94-109 (2016).
7. Truscott, J.: The case against grammar correction in L2 writing classes. Language learning 46(2), 327-369 (1996).

8. Chandler, J.: The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. Journal of Second Language Writing 12(3), 267-296 (2003).
9. Van der Kleij, Fabienne M., Feskens, Remco C. W., Eggen, Theo J. H. M.: Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes: A Meta-Analysis. Review of Educational Research 85(4), 475-511 (2015).
10. Ackerman, J., M.: Reading, writing, and knowing: the role of disciplinary knowledge in comprehension and composing. Research in the Teaching of English 25(2), 133-178 (1991).
11. Ito, F.: L2 reading–writing correlation in Japanese EFL high school students. The Language Teacher 35(5), 23-29 (2011).
12. Park, Y., Warschauer, M.: Syntactic enhancement and second language literacy: an experimental study. Language Learning & Technology 20(3), 180-199 (2016).
13. Kawamura, K., Kashiwagi, H., Kang, M.: An approach toward automatic error detection in learners' English writing based on the source language. In: 10th International Proceedings on Mobile, Hybrid, and On-line Learning, pp. 62-65. IARIA, Roma (2018).
14. Asencion Delaney, Y.: Investigating the reading-to-write construct. Journal of English for Academic Purposes 7, 140-150 (2008).
15. Kuhbandner, C., Pekrun, R.: Joint Effects of Emotion and Color on Memory. Emotion 13(2), 375-379 (2013).
16. Gali, N., Mariescu-Istodor, R., Hostettler, D., Fränti, P.: Framework for syntactic string similarity measures. Expert Systems with Applications 129, 169–185 (2019).
17. Klein, D., Manning, C., D.: Accurate unlexicalized parsing. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430. (2003).