ATLANTIS PRESS
Advances in Social Science, Education and Humanities Research (ASSEHR), volume 68
2017 International Conference on Advanced Technologies Enhancing Education (ICAT2E 2017)

Annotation of a Learner Corpus toward Development of an Error-cause Presenting Technique

Katsunori Kotani[1,a], Takehiko Yoshimi[2]

[1]School of Foreign Language, Kansai Gaidai University, 16-1 Nakamiya, Hirakata, Osaka, Japan;

[2]Department of Media Infomatics, Ryukoku University, 1-5 Yokotani, Oe-cho, Seta, Shiga, Japan.

[a]kkotani@kansaigaidai.ac.jp

Keywords: writing, learner corpus, annotation for article errors, feedback for error prevention

Abstract. Previous grammar checkers support learners' writing by indicating errors and correct usages. However, until the lack of linguistic knowledge that caused the errors is corrected, similar errors will recur. This study developed a learner corpus by annotating tags for the causes of article errors, and analyzed the distribution of error-cause tags with the aim of developing a technique to provide feedback on the causes of grammatical errors. The results suggested the necessity to extend error-cause tags, and the preferable extension conditions based on writing type and proficiency level.

Introduction

Feedback on grammatical errors is available with natural language processing techniques detecting errors in articles [1-4], which plays a tool role in technology-enhanced language learning [5]. Feedback on article errors is significant for English learners whose first language lacks an article system [6, 7] because such errors are substantial due to the frequency of article use [8]. One goal of feedback is for learners to acquire the lacking linguistic knowledge that caused the errors (henceforth, error causes) in order to prevent similar errors from recurring.

Given this goal, we intend to develop a technique that presents error causes as feedback. Development of this technique requires a learner corpus annotated with tags for error causes (henceforth, error-cause tags). We found no such learner corpus available, although most previous corpora included annotation information on the presence of errors and correct usages [9, 10].

Therefore, we developed and annotated a learner corpus with error-cause tags. We further examined the distribution of error-cause tags in order to validate the learner corpus as a language resource toward the development of an error-cause presenting technique. The results showed that the distribution was skewed: some error causes appeared frequently, while others infrequently, and the distribution depended on the type of writing and on the learners' proficiency levels.

Examination of Error-Cause Tags

Subjects and Learner Corpus. Subjects were university students who had more than 6 years of English as a foreign language learning experience in junior and senior high school. Their first language was Japanese. The subjects were classified into one of three groups based on their scores in the last 12 months on the Test of English for International Communication (beginner: 280 to 485; intermediate: 490 to 725; and advanced: 730 to 985), which is a popular English test in Japan. According to the Educational Testing Service, the mean score in Japan is 572.9 [standard deviation, (SD) = 174.4, range: 10-990]. The number of learners (n), mean score (m), and SD of the groups in this study were as follows: beginner group (n = 30, m = 404.7, SD = 51.3), intermediate group (n = 30, m = 632.5, SD = 69.8), and advanced group (n = 30, m = 864.3, SD = 69.0).

A learner corpus [11] was chosen for annotation of error-cause tags. This corpus covered two writing types: narratives (picture descriptions) and explanations (question answering). The writing data were compiled as follows: learners wrote sentences describing four pictures that illustrated a

series of events (at least five sentences per picture), and sentences answering 20 questions (one sentence per question was sufficient) about school, learning English and computer skills (e.g., "What are your favorite subjects?" and "How comfortable are you with using a computer?"). When describing the pictures, subjects needed to consider contextual information, such as using an indefinite article for a reference in the first picture, and a definite article for that reference in the following picture(s). They were prohibited from using dictionaries or any other reference books.

The writing data were already annotated with grammatical error tags on the basis of an annotation scheme [12]. This scheme targets lexical and grammatical errors classified into three types: replacement, redundant, and omission. These errors were evaluated by an English teacher.

The writing data consisted of 4,007 sentences (29,115 words). 639 instances of article errors appeared, and among them, 439 instances (68.7%) were errors due to the lack of articles.

Annotation of Error-cause Tags. Although there are three types of article errors (confusion, overuse, and lack), this study targeted only the lack of articles because this type of error is judged as more severe for English speakers to read [8]. Henceforth, the term "article error(s)" refers to errors in the lack of article use. All of the article errors (439 instances) were annotated with error-cause tags that described the correct use of articles [13, 14] as found in a grammar textbook for high-school students [15] in order for learners to comprehend the explanations of the error causes. The observed error cause tags were annotated according to a manually built decision tree (Figure 1). This annotation task was carried out by an English teacher who was also the author of this paper.

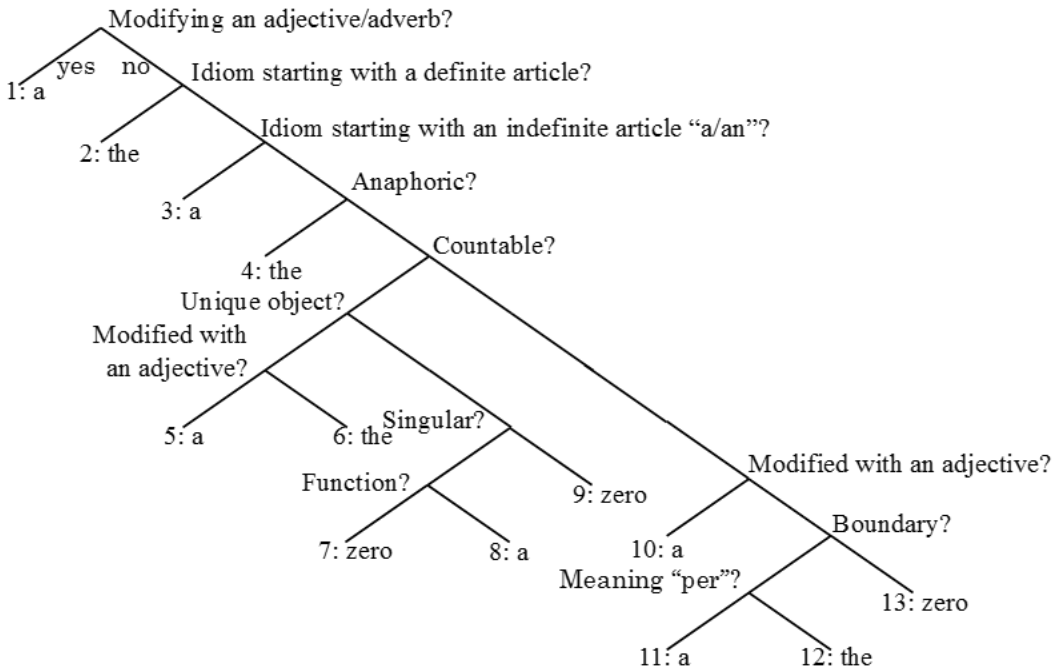

Fig. 1 Decision tree for the causes of article errors

The decision tree starts on the root node on which the user decides whether an article is used for an adjective/adverb such as "a little/few" or not. If the answer is "yes," the left-hand branch is taken. If the answer is "no," the right-hand branch is taken. Leaf nodes, e.g., "1: a" and "2: the" represent error-cause tags. The tree moves along the nodes for the idiomatic uses of (in)definite articles. The following nodes represent semantic properties of nouns, such as anaphoric and countable nouns. Note that the node for the semantic property of "function" refers to the use of "zero" article for a noun phrase whose functional meaning is focused on, as in "by bicycle" in contrast to "with a bicycle." Note also that the node for the semantic property of "boundary" refers to the use of "zero" article for an uncountable noun (e.g., "housework") that conveys the general meaning, not the specific meaning (e.g. "cleaning," "dishwashing," etc.).

Annotation information was represented following the error tags [12], as in "A woman said in <at crr="a"></at> loud voice, 'I love you!'" The error tag <at crr="a"></at> indicated that an error was related to the article "at," and that its correct form (crr) was the indefinite article "a." This annotation scheme was extended by adding error-cause tags with the numbers 1 to 13 corresponding to the leaf nodes on the decision tree of Figure 1 as follows: <at crr="a"></at_8>, where the number 8 refers to the leaf node "8: a" (the use of an indefinite article for a non-anaphoric singular noun).

Calculating the Distribution of Error-cause Tags. The raw frequency of the error-cause tags was counted for each leaf node from "1: a" to "13: zero." In order to confirm the influence of the writing type and the proficiency level of the learners, the frequency was standardized as the number of tags per 10,000 words ($\frac{raw\ frequency}{raw\ frequency\ for\ writing\ type\ or\ proficiency\ level} \times 10{,}000$).

Raw Frequency of Error-cause Tags. Table 1 shows the raw frequency of the error-cause tags. The error-cause tags "7: zero," "9: zero," and "13: zero" were excluded because they were deemed irrelevant to the target of this study. An extremely large number of errors were observed for the tags "4: the" (use of a definite article for an anaphoric referent) and "8: a" (use of an indefinite article for a singular non-anaphoric referent), which are considered basic article uses. Since the frequency of the error-cause tags was low, the corpus data should be extended for the development of an error-cause presenting technique.

Table 1 Distribution of error-cause tags by writing type and proficiency level

Error-cause Tag	Raw frequency	Narratives			Explanations		
		Beginner (4,097)	Intermediate (4,527)	Advanced (5,983)	Beginner (4,048)	Intermediate (4,801)	Advanced (5,659)
1: a	1	0	0	0	0	2	0
2: the	24	15	4	2	10	10	11
3: a	0	0	0	0	0	0	0
4: the	160	183	121	30	7	12	5
5: a	0	0	0	0	0	0	0
6: the	8	0	0	0	2	2	11
8: a	235	139	110	30	94	92	49
10: a	7	0	7	0	2	4	2
11: a	3	0	0	0	2	4	0
12: the	1	0	0	0	0	2	0
Total	439	337	243	62	119	129	78
		642			326		

Distribution of Error-cause Tags by Writing Type and Proficiency Level. Table 1 also shows the standardized frequency of error-cause tags by writing type and proficiency level. The total frequency of error-cause tags in the narratives was approximately twice as high as that of the explanations. This difference may be due to the conditions on the number of sentences. It is also supposed that the narratives had both contextually evoked and non-evoked referents, which led to the higher number of article errors. Given the difficulty of using of articles when describing pictures, the less proficient learners likely made more errors in the narratives than in the explanations. The beginner group made three times more article errors in the narratives than in the explanations. The difference was double for the intermediate group, and little difference was observed in the advanced group. The results show that corpus data should be modified through the use of picture descriptions.

The influence of English proficiency was observed in narratives for the error-cause tags "4: the" and "8: a." However, the influence of proficiency was not observed in explanations for the error-cause tag "2: the," which showed no difference among the groups, or for the error-cause tag "6: the," which showed the highest frequency in the advanced group. The examination of the error-cause tag

"6: the" in explanations by the advanced group showed that the article errors occurred due to the use of definite articles for superlatives such as "the most." Learners at the beginner and intermediate levels might not have used superlatives, which explains the low frequency of article-error-cause presenting tags regarding the error-cause tag "6: the." Hence, when adding corpus data for the development of an error-cause presenting technique, it is better to assign the condition that superlatives must be used to describe pictures or answer questions.

Summary

This study described an annotation scheme for error-cause tags that presents linguistic knowledge, the lack of which caused the errors, and reported the distribution of error-cause tags among beginner- to advanced-level English learners. The results suggest that our learner corpus needs to be extended by increasing error-cause tags to take the type of writing into consideration. However, this annotation scheme will help English learners learn how to use articles, and can also be applied to other grammatical errors such as the use of prepositions and auxiliaries.

References

[1] J. Lee, Automatic Article Restoration. Proc. of HLT-NAACL (2004) 31-36.

[2] R. Nagata, T. Wakana, F. Masui, A. Kawai, N. Isu, Detecting Article Errors based on the Mass Count Distinction. Proc. of IJCNLP (2005) 815-826.

[3] N.-R. Han, M. Chodorow, C. Leacock, Detecting Errors in English Article Usage with a Maximum Entropy Classifier Trained on a Large, Diverse Corpus. Proc. of LREC (2004) 1625-1628.

[4] G. Berend, V. Vincze, S. Zarriess, R. Farkas, LFG-based Features for Noun Number and Article Grammatical Errors. Proc. of CoNLL (2013) 62-67.

[5] R.P. Taylor, Introduction, in: R. P. Taylor (Ed.) The Computer in School: Tutor, Tool, Tutee. Teachers College Press, New York, NY, (1980), 1–10.

[6] R. Ellis, Y. Sheen, M. Murakami, H. Takashima, The Effects of Focused and Unfocused Written Corrective Feedback in an English as a Foreign Language Context. System 36(3) (2008) 353-371.

[7] E. Izumi, K. Uchimoto, T. Saiga, T. Supnithi, H. Isahara, Automatic Error Detection in the Japanese Learners' English Spoken Data. Proc. of ACL (2003), 145-148.

[8] K. Knight, I. Chander, Automated Postediting of Documents. Proc. of AAAI (1994) 779-784.

[9] A. Rozovskaya, D. Roth, Generating Confusion Sets for Context-sensitive Error Correction. Proc. of EMNLP (2010) 961-970.

[10] D. Dahlmeier, H.T. Ng, S.M. Wu, Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English. Proc. of NAACL Workshop (2013) 22-31.

[11] K. Kotani, T. Yoshimi, Analysis of the Association between Reading Proficiency and Frequency of Writing Errors Using an Error-tagged Learner Corpus for Multiple Linguistic Skills. ICIC Express Letters: An Intl. J. of Research and Surveys (2016) 569-573.

[12] E. Izumi, K. Uchimoto, H. Isahara, Error Annotation for Corpus of Japanese Learner English. Proc. of Workshop on Linguistically Interpreted Corpora (2005) 71-80.

[13] A.J. Thomshon, A.V. Martinet, A Practical English Grammar, third ed., OUP, Oxford, (1985).

[14] R. Quirk, S. Greenbaum, G. Leech, J. Svartvik, A Comprehensive Grammar of the English Language. Longman, London, (1985).

[15] A. Ishiguro (Ed.) Sogo Eigo (Comprehensive English) Forest, Kirihara Shoten, Tokyo, (2009).