Grammatical Error Annotation for Korean Learners of Spoken English

Hongsuck Seo[1)], Kyusong Lee[1)], Gary Geunbae Lee[1)], Soo-Ok Kweon[2)], Hae-Ri Kim[3)]
Department of Computer Science and Engineering, POSTECH, Korea[1)]
Division of Humanities and Social Sciences, POSTECH, Korea[2)]
Department of English Education, Seoul National University of Education, Korea[3)]
E-mail: {hsseo, kyusonglee, gblee, sook}@postech.ac.kr, hrkim@snue.ac.kr

Abstract

The goal of our research is to build a grammatical error-tagged corpus for Korean learners of Spoken English dubbed Postech Learner Corpus. We collected raw story-telling speech from Korean university students. Transcription and annotation using the Cambridge Learner Corpus tagset were performed by six Korean annotators fluent in English. For the annotation of the corpus, we developed an annotation tool and a validation tool. After comparing human annotation with machine-recommended error tags, unmatched errors were rechecked by a native annotator. We observed different characteristics between the spoken language corpus built in this study and an existing written language corpus

Keywords: grammatical error, Korean, spoken corpus

1. Introduction

Recently, language learning has drawn a significant attention in the field and consequently, computer-assisted language learning (CALL) has been intensely researched to reduce the cost of language teaching and learning. One research topic related to CALL is grammatical error detection and correction. While many existing CALL systems helping learners develop their grammar skill have used hand-crafted and pre-scheduled materials, automatic methods have been sought for the development of learners' grammar skill. Since knowledge of language grammar is needed in some way to detect or correct grammatical errors, several resources have been used: learner corpora, artificial error corpora, hand-crafted parsing rules including grammatical errors, etc. (Nicholls, 2003; Granger, 2004; Izumi, 2004; Lee, 2011; Schneider, 1998). Among these different resources, learner corpora, which are a set of raw text or speech tagged with grammatical error types, and sometimes corrections, are used for a number of purposes such as error analysis and the influence of the learner's mother tongue on errors. They are also extremely useful for data-driven approaches to error detection and correction.

In the present study, we built a spoken language corpus of Korean learners of English tagged with an existing error tagset. We call the corpus the Postech Learner Corpus (POLC) and it is available[1]. Although there are several corpora available, this research is meaningful for the following reasons. Firstly, considering that the Japanese Learner English (JLE) corpus is, to the best of our knowledge, the only fully tagged L2 corpus for spoken English, our corpus will compensate the clear lack of data in the available learner speech to date. Secondly, because the JLE corpus is constructed for Japanese speakers, it is necessary to build a new dataset for Korean learners in order to analyze errors and develop automatic error detection and correction systems of these learners. Finally, the tasks learners performed in collecting data for corpus construction were different: the JLE corpus contains interviews between learners and interviewers, however, our corpus, POLC, consists of contextualized story-telling tasks based on the picture description by Korean learners of English.

This paper is organized as follows. In section 2, we will show the overall method of building the corpus including the learners' task, tagset and tools used. In section 3, we analyze the characteristics of the built dataset, comparing with a written language corpus for Korean learners. Lastly we summarize our work and outline future plans.

2. Method

In this section, we show how we developed the POLC. A total of 42 learners participated in data collection and, each student was asked to describe five different picture books with ranging 10-12 pages. Each speech data was collected and transcribed, then six Korean speakers who were fluent in English, annotated errors. Each annotator was given 35 speech data generated by seven learners. We used two tools for the development process: an annotation tool and a validation tool. The validation tool was used to reconfirm annotated tags.

2.1 Learners' Task

We collected raw speech data from 42 Korean learners of English who were the university students with various majors. The participants' task was to describe each of five picture books containing interesting stories for young adults. For the purpose of the experiment, we eliminated the letters and presented pictures only to make the learners guess the flow of the story and describe it in speech. Participants saw the pictures on the computer screen page by page and were asked to describe each book in three minutes, each student providing five three-minute descriptions. Each description was recorded and carefully transcribed without changing any of learners' original errors.

[1] http://isoft.postech.ac.kr

Code	Description
F	wrong Form used
M	something Missing
R	word or phrase needs Replacing
U	word or phrase is Unnecessary
D	word is wrongly Derived
I	word is wrongly Inflected

Table 1: Error codes in CLC tagset for each error types

Code	Description
A	Pronoun (Anaphoric)
C	Conjunction (linking word)
D	Determiner
J	Adjective
N	Noun
Q	Quantifier
T	Preposition
V	Verb (includes modals)
Y	Adverb (-lY)
P	Punctuation

Table 2: Sub-codes in the CLC tagset for each word

Code	Description
AS	incorrect Argument Structure
AGA	Anaphoric (pronoun) AGreement error
AGD	Determiner AGreement error
AGN	Noun AGreement error
AGQ	Quantifier AGreement error
AGV	Verb AGreement error
CD	wrong Determiner because of noun Countability
CE	Compound Error
CN	Countability of Noun error
CQ	wrong Quantifier because of noun Countability
ID	IDiom error
L	inappropriate register (Label)
S	Spelling error
SA	American Spelling
SX	Spelling confusion error
TV	wrong Tense of Verb
W	incorrect Word order
X	incorrect formation of negative

Table 3: Additional codes of the CLC tagset for exceptions

2.2 Tagset

We used the Cambridge Learner Corpus (CLC) tagset[2] to annotate the transcribed learner speech. We chose the CLC tagset, an existing error tagset for written language English, rather than the JLE tagset, which is designed for spoken English, for two reasons. First, the error tagset on written English includes all the errors in spoken English as well. Second, in the future research, we are planning to develop a grammatical error detection and correction system and extend it to cover not only spoken errors but also written errors.

The structure of the CLC tagset is mainly a combination of error types and word classes, as shown in Table 1 and Table 2, respectively. For example, an RV tag is a combination of the error type R, replacement error, and the word class V, verb, representing a verb replacement error as in the following example:

The learner sentence: *What do you believe about that?*
The tagged sentence:

What do you
<NS type="**RV**">*believe*<c>*think*</c></NS>
about that?

The errors are enclosed by a tag <NS> with their corresponding error type. The tag denotes an incorrect word and <c> denotes its corresponding correction. In cases where we cannot find any specific word class for an error, we simply use one letter tag, for instance, U, for unnecessary errors. The CLC tagset also includes some special tags for exceptional cases (Table 3)[3].

2.3 Annotation Tool

We developed an annotation tool for grammatical errors. The basic function of the tool is to help the annotators unfamiliar with the XML format annotate tags. The annotators simply type the sentence with one error correction to the textbox and choose the type of the corrected error from the combo-box (Figure 1). After reviewing faulty passages, Annotators input whole sentences with the errors corrected into the tool. The tool than compares the corrected sentence to the original text and automatically generates an XML structured tags. Annotators were also provided with examples of tags to assist their work.

[2] http://www.cup.cam.ac.uk/gb/elt/catalogue/subject/custom/item3646603/Cambridge-English-Corpus-Cambridge-Learner-Corpus/?site_locale=en_GB

[3] The error type SA is included in the tagset because the purpose was for learners to learn British English in the CLC. However, the learners in our task do not specifically focus on learning British English so the SA error is excluded in this work.

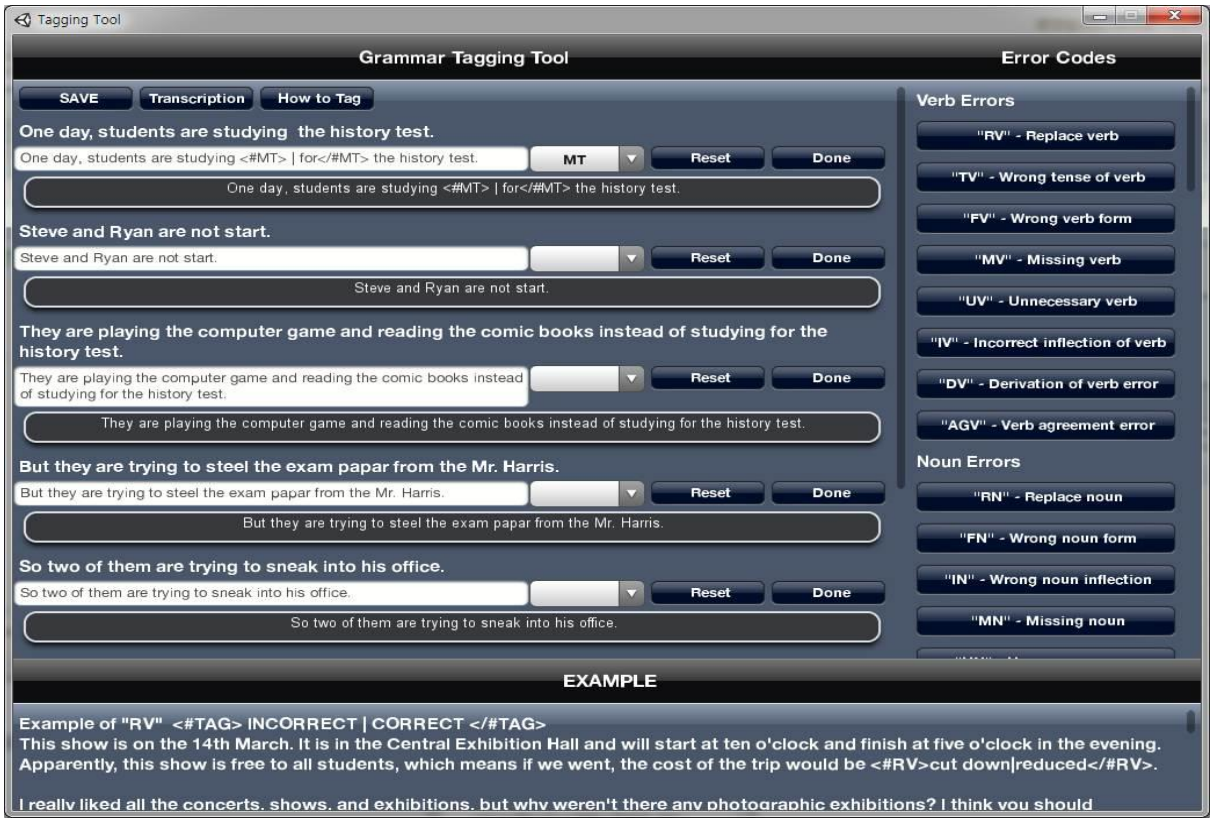

Figure 1: Screen shot of the annotation tool

Features	Accuracy
Word	0.787
Word+POS	0.830
Word+POS+Lemma	0.866
Word+POS+Lemma+POS equality	0.870

Table 4: Accuracy of the validation

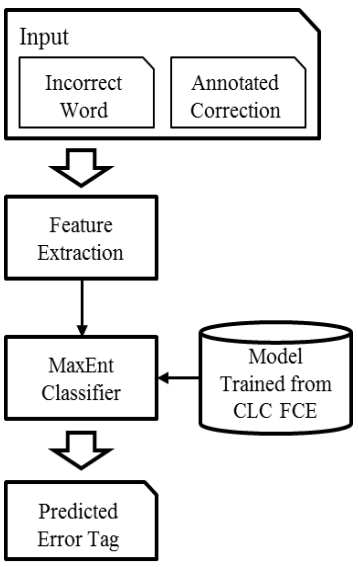

Figure 2: The overall architecture of the validation tool

2.4 Validation

Although the annotators who transcribed the speech and found grammatical errors were fluent in English, they could not reliably distinguish with perfect precision. Thus, it is necessary to employ a validation tool to increase the quality of corpus. The validation tool could predict the correct tag type when given the learner's incorrect form and the annotator's correct form. Only for the annotated tag that did not match to the tool-generated tag did a native speaker annotator recheck to increase reliability. We developed the validation tool using Maximum Entropy (MaxEnt) classification technique (Figure 2). The CLC FCE[4] corpus (the collection of First Certificate in English exams in the CLC) is used for training model. We employed some linguistic features, such as words, lemmas, and Part-of-Speech (POS) of incorrect words and its correction. We used 80 % of the CLC FCE corpus for training MaxEnt and the rest 20 % for test. The result shows the high accuracy with the features: words, POS, lemma, and POS equality (Table 4). As the model for the validation tool, we used the full set of the CLC FCE corpus.

3. Data Analysis

3.1 Validation Test

A total of 67 % of annotations were classified as valid tags and 33 % were unmatched tags with the trained model.

[4] http://www.ilexir.com/181

CLC FCE [36841 words/written]			POLC [22423 words/spoken]		
Tag	Count	Rate	Tag	Count	Rate
MD	324	12.28	TV	581	15.40
S	201	7.62	AGV	528	13.99
RT	163	6.18	MD	275	7.29
RV	140	5.31	RV	262	6.94
RP	135	5.12	FV	217	5.75
TV	118	4.47	UD	195	5.17
MT	118	4.47	RN	192	5.09
MP	102	3.87	MT	152	4.03
UD	90	3.41	UC	115	3.05
UT	88	3.33	FN	91	2.41

Table 5: Error types with the 10 highest frequencies in the Korean products of the CLC FCE corpus and the POLC

Since most errors made by both the validation tool and human annotators were caused by the ambiguity of some error types, these errors were usually filtered as unmatched tags. The unmatched tags were rechecked by a qualified native speaker annotator.

3.2 Comparison to Written Language Corpus

In this section, we conduct a contrastive analysis between the error distributions of spoken language and written language corpora of Korean learners of English, using the POLC and the CLC FCE. For the analysis, we extracted only the essays written by Korean learners from the CLC FCE. Considering error characteristics, we divided the errors into two groups: correctable errors and uncorrectable errors. Correctable errors, such as AGV and FV errors, are the errors learners can correct given some time, whereas uncorrectable errors including MD and RV errors are not. While correctable errors are deterministic and usually morphological or simple structural errors, uncorrectable errors involve verb semantic errors which may require native speaker's intuition. In the written language corpus the errors occurring with the highest frequencies were mostly uncorrectable errors. In the spoken language corpus, however, correctable errors occurred frequently (Table 5). This phenomenon is because of the fundamental differences between the written and spoken languages. Since learners can take some time to check their product during written tasks, those correctable errors can be corrected after rechecking. When it comes to spoken tasks, learners cannot take any time to recheck and correct their product because spoken tasks are online tasks. TV errors, the most highly frequent error type in spoken language corpus, include not only tense but aspect and voice of a verb, which indicates that some tense errors are correctable but the others are not. This may explain the increased error rate of TV in spoken English compared to written English.

4. Conclusion

In this work, we developed a spoken English corpus for Korean learners. After collecting raw story telling speech from Korean learners, six annotators fluent in English transcribed and annotated using the CLC tagset. During the annotation process, the annotators used an annotation tool which is designed for people without any knowledge of XML. We also developed and used a validation tool which predicts the error tags of given corrections to raise the quality of the corpus. The unmatched annotations with the validation tool were rechecked by a qualified native speaker annotator. Our follow-up research is to develop an automatic error detection and correction system with the POLC. We will also extend the research to written language corpus of Korean learner English.

5. Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0027953)

6. References

Granger, S. (2004). Computer learner corpus research: current status and future prospects. *Applied corpus Linguistics: A Multidimensional Perspective.* Amsterdam & Atlanta: Rodopi, pp. 123—145.

Izumi, E.; Uchimoto, K. and Isahara, H. (2004). The overview of the sst speech corpus of Japanese Learner English and evaluation through the experiment of automatic detection of learners' errors. In *Proceedings of Language Resource and Evaluation Conference (LREC)*, Lisbon, Portugal, pp. 1435—1438.

Lee, S.; Noh, H.; Lee, K.; Lee, G. G. (2011). Grammatical error detection for corrective feedback provision in oral conversations. In Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, San Francisco, CA, pp. 797—802.

Nicholls, D. (2003). The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 Conference (CL 2003)*, Lancaster, UK, pp . 572—581.

Schneider, D. and McCoy, K. (1998). Recognizing syntactic errors in the writing of second language learners. In proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL) and the 17th International conference on Computational Linguistics (COLING), Montreal, Canada, pp. 1198—1204.