Computer Assisted Language Instruction Consortium

CALICO JOURNAL (ONLINE) ISSN 2056–9017

Article

Automated Error Detection for Developing Grammar Proficiency of ESL Learners

Hui-Hsien Feng, Aysel Saricaoglu, and Evgeny Chukharev-Hudilainen

Abstract

Thanks to natural language processing technologies, computer programs are actively being used not only for holistic scoring, but also for formative evaluation of writing. CyWrite is one such program that is under development. The program is built upon Second Language Acquisition theories and aims to assist ESL learners in higher education by providing them with effective formative feedback to facilitate autonomous learning and improvement of their writing skills. In this study, we focus on CyWrite's capacity to detect grammatical errors in student writing. We specifically report on (1) computational and pedagogical approaches to the development of the tool in terms of students' grammatical accuracy, and (2) the performance of our grammatical analyzer. We evaluated the performance of CyWrite on a corpus of essays written by ESL undergraduate students with regards to four types of grammatical errors: quantifiers, subject-verb agreement, articles, and run-on sentences. We compared CyWrite's performance at detecting these errors to the performance of a well-known commercially available AWE tool, Criterion. Our findings demonstrated better performance metrics of our tool as compared to Criterion, and a deeper analysis of false positives and false negatives shed light on how CyWrite's performance can be improved.

KEYWORDS: AUTOMATED WRITING EVALUATION; DESIGN-BASED RESEARCH; ERROR CORRECTION; EVALUATION OF AWE SYSTEMS; FORMATIVE FEEDBACK; GRAMMATICAL ERRORS

Affiliation

Iowa State University.
email: hhfeng@iastate.edu (corresponding author)

CALICO JOURNAL VOL 33.1 2016 49–70

doi : 10.1558/cj.v33i1.26507

equinox www.equinoxpub.com

equinoxonline

1. Introduction

The role of explicit teaching of grammar has changed in the field of second language acquisition (SLA) throughout its history, alongside changes in theoretical views of language. While language teaching meant the teaching of grammar in the 1850s, grammar was eliminated from language classes until the late 1900s (Nassaji & Fotos, 2011; Richards & Rodgers, 2001). With functional perspectives of grammar gaining momentum, teaching grammar explicitly bacame important again in the 1970s and 1980s (Nassaji & Fotos, 2011). Since then, the focus on grammar has drawn the attention of teachers and researchers to learners' grammatical errors and ways of treating them.

Empirical studies suggest that language learners' grammatical errors should be addressed through corrective feedback to help improve their grammatical accuracy (Ferris & Roberts, 2001). Furthermore, Bitchener and Ferris (2012) and Heift (2004) argue that feedback about grammar should incorporate detailed metalinguistic explanations about the rules that the learner violates when making grammatical errors. From the practical viewpoint, providing grammatical feedback to learners is problematic for two reasons. First, locating and marking errors is a very time-consuming task for instructors. This time burden increases if the instructor needs to identify and explain specific grammatical rules to the learner while providing corrective feedback. Second, providing accurate grammatical feedback requires working knowledge of descriptive grammar and comparative linguistic typology, which some teachers might lack if they have not received sufficient training in linguistics (Ferris, 2010).

Over the last few years, a growing body of research in computer-assisted language learning (CALL) has focused on responding to the challenge of providing accurate grammatical feedback to learners through the development and use of automated writing evaluation (AWE) tools based on natural language processing (NLP) and machine-learning technologies. These tools, such as Criterion, the web-based application developed by Educational Testing Service, were originally designed to support grading of the essay components of large-scale standardized tests, but they have been repackaged to provide feedback on writing for learners (Burstein, 2003). For instance, Criterion evaluates learners' texts in terms of five categories: grammar, usage, mechanics, organization and development, and style. The types of grammatical errors that Criterion detects and provides feedback on include sentence fragments, missing commas, run-on sentences, garbled sentences, subject-verb agreement errors, ill-formed verbs, pronoun errors, possessive errors, and wrong or missing words (Attali, 2004; Hagerman, 2011; Shutler, 2012)

Classroom-based research has confirmed the potential of Criterion in teaching L2 grammar, but has also found issues to be addressed, including

the accuracy, clarity, and explicitness of formative feedback; learners' and teachers' confidence in AWE scoring; questions of when and how learners should use AWE while learning to write academically; and the usability and functionality of the system itself (Li, Feng, & Saricaoglu, in press; Li, Link, & Hegelheimer, 2015; Li, Link, Ma, Yang, & Hegelheimer, 2014; Link, Dursun, Karakaya, & Hegelheimer, 2014; Ranalli, Chukharev-Hudilainen, & Link, 2014). These issues make it clear that current AWE tools have weaknesses and they need to be improved for better learning and teaching practices and outcomes. ESL learners and instructors are in need of better performing AWE tools that they can use in academic writing settings. To address this need, the authors of the present paper developed a novel and customizable AWE system called CyWrite. In contrast to existing AWE tools originally designed for standardized testing, CyWrite was built to support not only testing but also the teaching and learning of L2 writing and research (Chukharev-Hudilainen & Saricaoglu, 2014). CyWrite relies on a hybrid (statistical and rule-based) NLP framework to detect various word-, sentence-, paragraph-, and text-level features, for example, spelling errors, problematic stylistic choices, certain discourse patterns (Chukharev-Hudilainen & Saricaoglu, 2014), and, of course, grammatical errors. Formative feedback is generated based on the detected features and delivered either concurrently with the composition process (e.g., in the form of a red squiggly line, for spelling errors, or comments on the margin, for grammatical errors) or episodically, that is, after students complete a writing session and submit their draft for automated evaluation. In the present paper, we report on the development and evaluation of the linguistic feature detection capabilities in CyWrite, specifically focusing on the identification and classification of certain grammatical errors in learners' texts. Feature detection is the first important step that leads to the provision of automated feedback, and the accuracy of detection is crucial for the quality of feedback provided (Chapelle, Cotos, & Lee, 2015).

The research and development of the CyWrite system has been guided by the design-based research (DBR) paradigm, which is a 'paradigm for the study of learning in context through the systematic design and study of instructional strategies and tools' (The Design-Based Research Collective, 2003: 5). DBR 'posits synergy between practice and research in everyday settings' (Wang & Hannafin, 2005: 13). The DBR approach underlines the importance of documenting the entire process of the development, evaluation, and revision of a pedagogical intervention or learning tool. Unlike more traditional predictive research, which aims at establishing statistically significant bases for rejecting predefined hypotheses by conducting controlled experiments, DBR focuses on continuous, iterative refinement of interventions, which leads to the formulation of design principles ('best practices') for the development and

enactment of future, improved versions of such interventions (Kennedy-Clark, 2013; Plomp, 2009). Although the DBR approach has been applied in educational technology studies for over a decade (Anderson & Shattuck, 2012; Kelly, Lesh, & Baek, 2008), it was not until recently that it started to gain momentum as a methodological framework for research in CALL (e.g., Echevarria, Short, & Powers, 2006; Lund, 2005, 2008; Lund & Smordal, 2006; Hung, 2011; Pardo-Ballester & Rodríguez, 2009, 2010, 2013; Yutdhana, 2005a, 2005b). The CALICO monograph issue *Design-based Research in CALL* edited by Rodriguez and Pardo-Ballester (2013) collects a number of seminal papers that provide more details on CALL research based on the DBR approach.

Within DBR, research and implementation of a pedagogical intervention or a learning tool is performed in several iterations, and the data yielded by a previous iteration is used to inform further refinement of the tool, the intervention, or both. This paper describes the initial iteration in the development and research of CyWrite's grammatical error detection engine. We note that this initial iteration does not involve using the system in the actual classroom. Students' interaction with the ultimate CALL tool that will be developed in the future is only simulated at this stage of development by conducting analyses of a learner corpus. This iteration was guided by the following two research questions (RQs):

1. How well does CyWrite detect grammatical errors in ESL students' academic writing?
2. If CyWrite's performance is not perfect, what are the reasons for that and how the performance can be improved?

The rest of this paper is organized as follows. We begin with the description of our methodology in terms of the NLP approach selected for the development of CyWrite, and in terms of the methods used to evaluate the performance of CyWrite and compare it to the baseline established by Criterion. Finally, we will present our results, discuss pedagogical implications, and suggest directions for future work.

2. Method

2.1. Development

The first prototype of the CyWrite's grammatical error detection module is designed as a hybrid NLP system that identifies features of interest in the input text by combining statistical parsing with manually developed rules. During the first (statistical) stage of analysis, Stanford CoreNLP (Klein & Manning, 2003) is used to split the input text into sentences and word tokens, apply part-of-speech (POS) tags, create a constituents tree that

represents the grammatical structure of the sentence, and identify Stanford Typed Dependencies between words in the sentence (De Marneffe, MacCartney, & Manning, 2006) (see Figure 1). Despite the common-sense assumption that automatic parsing may not be reliable in the case of language learners' writing, Stanford CoreNLP has been found to perform reasonably well even on grammatically imperfect texts produced by language learners (O'Donnell, 2008). Similarly, in their system which provides feedback to learners based on the complexity and diversity of cause-and-effect expressions in their essays, Chukharev-Hudilainen and Saricaoglu (2014) used Stanford CoreNLP and found that parser errors were responsible for as little as 20% of cases when cause-and-effect expressions were incorrectly identified.

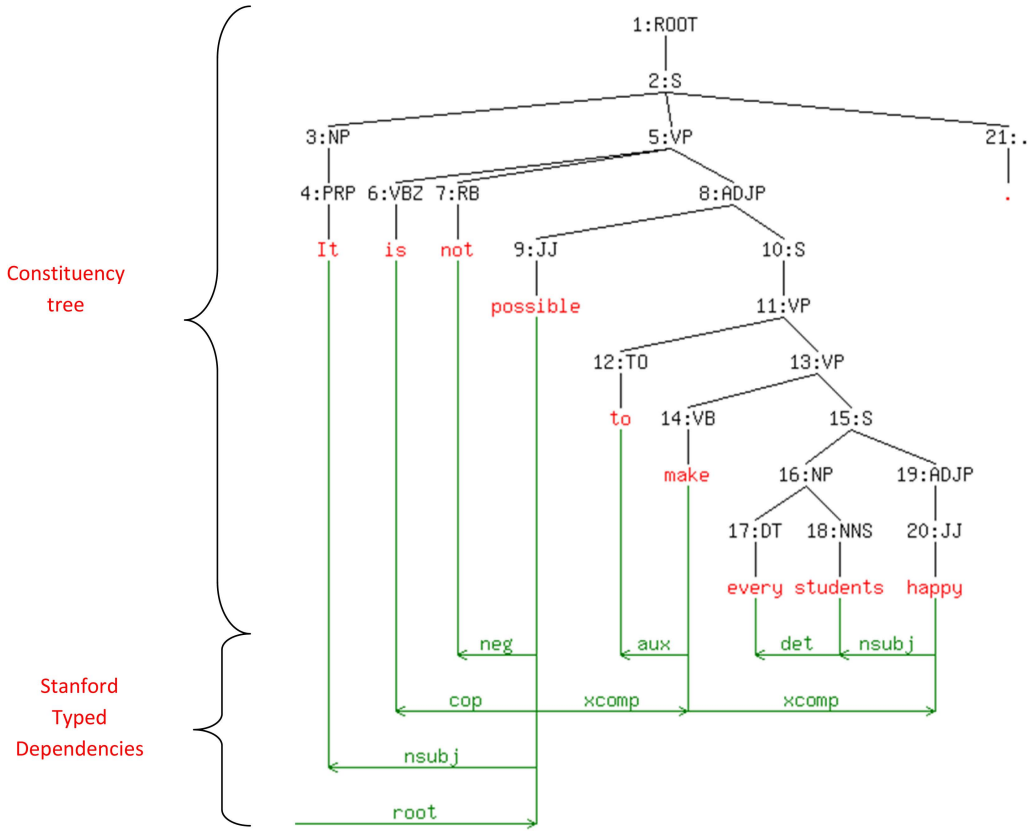

Figure 1: Parse tree of a sentence with a quantifier error

Stanford CoreNLP produces a formal description of the linguistic structures identified in the sentences. This description is produced as a document in the Extensible Markup Language (XML), which serves as the input for the second (rule-based) stage of analysis. This stage is handled by the software developed by the authors of the present article. During this stage, the XML

document is automatically converted into a Prolog program. Prolog is an all-purpose logical programming language with its roots in first-order logic (Bramer, 2013). For every grammatical error investigated in the present paper, we wrote a set of rules in Prolog to identify irregularities in the constituent tree structure corresponding to the error in question. This approach is known as mal-rules (Bender, Flickinger, Oepen, Walsh, & Baldwin, 2004; Leacock, Chodorow, Gamon, & Tetreault, 2010; Meurers, 2012; Schneider & McCoy, 1998). For the iteration reported in the present paper, we focused on four categories of grammatical errors: quantifier errors, subject-verb agreement errors, article errors, and run-on sentence errors. The error types and examples of errors are presented in Table 1.

Table 1: Error Types in CyWrite and Criterion and Examples of Errors

Target grammatical error in CyWrite	Equivalent grammatical error in Criterion	Examples of errors
Quantifiers	Determiner Noun Agreement	There were **four leader** in Mecca at that time.
Subject-Verb agreement	Subject-Verb agreement	**He work** hard on his study.
Articles	Missing or Extra Article; Wrong Article	He always got **good result** in piano performances.
Run-on sentence	Run-on sentence	Scientists are trying to create new invention to avoid the situation I mentioned before**,** they may come up with their ideas in recent years.

The four error types that we report in this study have been given priority since they are frequent errors made by learners (Ferris, 2006). The frequency of errors as identified by Criterion in our learner corpus is given in Figure 2, which also shows that these four error types are commonly committed by undergraduate ESL students.

In the process of creating the mal-rules, we depended on a number of sources such as textbooks used in ESL classrooms (e.g., *The Everyday Writer* (Lunsford, 2012)) to identify the grammatical patterns for an error category; online corpora, such as the Corpus of Contemporary American English (COCA) (Davies, 2010), to find examples with the target grammatical structure; online dictionaries to retrieve and verify lexicogrammatical information on the target words, phrases, or larger structures; and our own corpora of both professional and learner texts to test our rules in the process of development. We should make two clarifying notes. First, when developing rules for articles, we only focused on those cases that are based on: (a) the countability of the nouns; and (b) the initial sounds of nouns that they

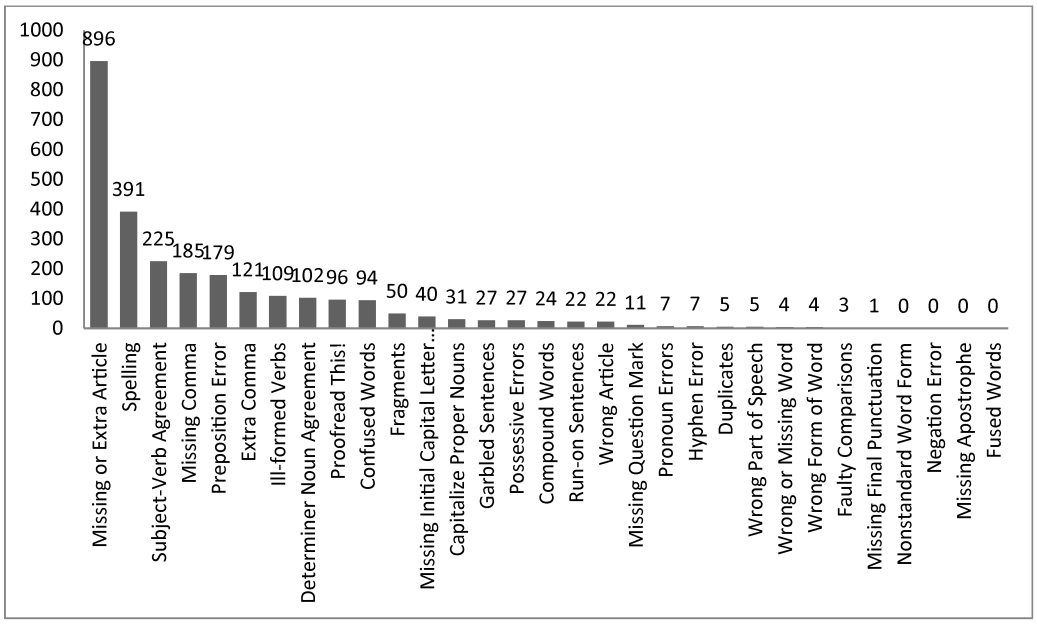

Figure 2: Error frequency retrieved from Criterion

modify. Other frequent error types, such as the incorrect use of *the* instead of *a/an* with singular count nouns, are beyond the scope of our AWE tool, especially because they are barely treatable and largely subjective. For example, our attempted annotation of such errors in our learner corpus failed with very low inter-rater reliability, although the annotation was done by experienced ESL teachers and involved repeated calibration. Therefore, we have left these types of article errors for future work.

Second, we should give a definition of a run-on sentence: 'a sentence containing two or more clauses not connected by the correct conjunction or punctuation' (Run-on sentence, n.d.). This error has been found frequently in both native (Connors & Lunsford, 1988; Lunsford & Lunsford, 2008) and nonnative (Hinkel, 2011; Sun, 2014; Wu & Garza, 2014) speakers' writing.

2.2. Evaluation

To aid in the development and evaluation of CyWrite, a database of essay drafts composed by students enrolled in a university course in academic writing for nonnative speakers of English was used. This database is maintained by the institution's ESL program for curricular purposes. The drafts in the database were written by intermediate-high and advanced-low level ESL students. Proficiency levels were determined based on the descriptors in the American Council on the Teaching of Foreign Languages (ACTFL). The essay prompts varied, including *Adapting to new technology, Breaking traditions, Career decisions, Choosing a job, Data Mining, Peer pressure, Computer and*

privilege, Cultural analysis, Summary of a book/movie, Role model, and *Discussion of Global Economics*. Using a computer program developed by one of the authors, 5,500 sentences were randomly sampled from the database along with corrective feedback generated by Criterion.

The entire corpus was manually annotated for all four error types. Each error type was treated as a separate annotation task. For each sentence in the corpus, the annotators had three options: (1) there is no error of this type in the sentence; (2) not sure; (3) the error of this type is present in the sentence. During calibration, all annotators were assigned the same sentences to work on. The reliability of annotation was assessed by computing Krippendorff's α (Hayes & Krippendorff, 2007; Krippendorff, 2011) with the interval metric. The reliability of interval $\alpha \geq 0.75$ was set as a target for calibration. If the actual reliability was substantially below the target, annotators were presented with a list of their disagreements and had an opportunity to discuss their decisions and improve their calibration. Once a satisfactory level of reliability was achieved, each of the annotators was assigned a random subset of the remaining corpus. At least 20% of the sentences were randomly selected to serve as the reliability subset and assigned to two annotators. All authors of this paper served as annotators, assisted by other CyWrite team members who contributed to the annotation process. We note that this practice of assigning only a subset of sentences to two annotators for reliability purposes saved a lot of effort, but prevented us from resolving each annotation discrepancy that might have arisen outside the reliability subset. That is, for up to 80% of the sentences, only one annotation was done, and if the annotator made a mistake, there was no way to identify or correct it. On the other hand, the reliability subset (i.e., sentences which were assigned to at least two annotators) was selected randomly, and the annotators did not know which sentences belonged to the reliability subset and which did not. As a result, we were able to extrapolate the reliability measures to the entire corpus; Krippendorff's α also explicitly supports incomplete (missing) annotation data.

Upon annotation, the corpus was randomly split into two parts: 70% of the sentences were designated as the development corpus and were used to assess the performance of CyWrite while developing and improving the rules, and the remaining 30% were set aside for the final evaluation of CyWrite's performance. Because each sentence in the corpus was associated with Criterion's grammatical feedback, CyWrite's performance was evaluated both in relation to the gold standard established by the manual annotation, and in comparison to the AWE baseline provided by Criterion.

Precision (P) of automatic error detection was evaluated as the number of correctly identified errors divided by the total number of errors identified by the tool. Therefore, when the tool produced false alarms, also called

false positives (i.e., it incorrectly flagged grammatical errors in sentences that were in fact grammatically correct), the value of P decreased. Recall (R) was calculated as the number of errors that the tool was able to find divided by the total number of errors in the corpus. R went down when the tool missed certain errors found by the annotators (these instances are also referred to as false negatives). A derived measure called the F-score, the harmonic mean of P and R: $F = 2 \times P \times R / (P + R)$. Accuracy was calculated as a simple percentage of sentences correctly classified as erroneous or error-free by the tool.

2.3. Analysis of sources of CyWrite Failures

After the calculation of precision, recall, F-score, and accuracy was done, the text files of false positives and negatives were generated from CyWrite, and manual analyses were conducted by the authors to categorize the sources of CyWrite failures. The categories used for this analysis are parser failures, rule issues, annotation mistakes, learner language problems, and others (Chukharev-Hudilainen & Saricaoglu, 2014).

3. Results

We follow Leacock *et al.*'s (2010) suggestions for improving consistency of reporting AWE evaluation results, which includes: (1) reporting of precision and recall in addition to the F-score; (2) having more than one annotator, and reporting the reliability between annotators; (3) providing detailed information about the annotated corpus and evaluation materials; (4) specifying the targets of the grammatical error detection; and (5) articulating the inclusion or exclusion of certain usages in or from an error type. In our previous section, we reported the last two items suggested by Leacock *et al.* (2010); we will report the first three items in this section.

3.1. Performance of the CyWrite grammar analyzer

We evaluated the performance of CyWrite by comparing its error detection in four grammar error categories both to the performance of Criterion and to the manual detection of the errors by human annotators.

The results for CyWrite error detection performance are presented in Table 2. Regarding quantifier error detection, Table 2 demonstrates that 75% ($P = 0.75$) of CyWrite's detection of quantifier errors was correct. CyWrite performed better than Criterion, which extracted 65% ($P = 0.65$) of the quantifier errors correctly. Compared to all the quantifier errors that were detected manually by the annotators, CyWrite could only detect 63%, which was still good in comparison with Criterion's recall estimate (37%). While CyWrite demonstrated a better performance of detecting quantifier errors than Criterion, the

accuracy of the two tools was found to be similar. Overall, the F-scores showed that CyWrite quantifier error detection performed better than Criterion.

As for CyWrite's performance at detecting subject-verb agreement errors, Table 3 shows that only 61% of CyWrite's detection (P = 0.61) was correct, which is lower than Criterion's precision (P = 0.80). However, CyWrite detected more errors (R = 0.67) than Criterion did (R = 0.49). In terms of

Table 2: Evaluation Results of CyWrite Error Detection

	Reliability of Annotation		Performance of CyWrite				Performance of Criterion[1]			
Error Type	PA	α	P	R	F	Accuracy	P	R	F	Accuracy
Quantifiers	0.98	0.74	**0.75**	0.63	**0.68**	**0.98**	0.65	0.37	0.47	0.97
Subject-Verb Agreement	0.95	0.71	0.61	**0.67**	**0.64**	0.94	**0.80**	0.49	0.61	**0.95**
Articles	0.92	0.74	0.56	**0.76**	**0.65**	0.84	**0.64**	0.55	0.59	0.84
Run-on Sentence	0.97	0.75	0.62	**0.30**	**0.41**	**0.95**	**0.80**	0.04	0.07	0.94

Note. PA – simple percent agreement without correction for chance, α – Krippendorff's Alpha reliability, P – precision, R – recall, F – F-score. Boldface number show superior performance.

accuracy, Criterion was slightly more accurate (95%) than CyWrite (94%). However, on the whole, CyWrite's F-score (F = 0.64) was higher than Criterion's (F = 0.61).

With respect to the CyWrite's detection of article errors, its accuracy was the same as Criterion's (84%). Although it extracted 56% article errors correctly (P = 56%), it was lower than Criterion's performance (64%). However, its recall (76%) was higher than Criterion's (55%), which shows that CyWrite found more article errors which were also flagged by human annotators than Criterion. Similarly, CyWrite reached higher recall, F-score, and accuracy than Criterion did in terms of the detection of run-on sentence errors.

3.2. Analysis of CyWrite's failures

Although the performance of CyWrite was found to be comparable, and in many cases superior, to the industry baseline established by Criterion, we analyzed the imperfections of CyWrite's grammatical error detection, that is, its false positives and false negatives (see Table 3).

Table 3: Types of CyWrite's Failures

	Correct hits	False positives	False negatives
Quantifiers	38	13	22
Subject-verb agreement	82	53	40
Articles	232	182	72
Run-on sentence	31	19	71

As seen in Table 3, CyWrite had false positives and false negatives in all error types. False positives and false negatives are very important since understanding their sources is the key of improving CyWrite rules for a better performance of error detection. In the following section, we elaborate on the causes of CyWrite's false positives and false negatives in quantifiers, subject-verb agreement, articles, and run-on sentence error detections.

Issues with quantifier error detection

The analysis of CyWrite's performance at detecting quantifier errors yielded 13 false positives. This means that CyWrite raised false alarms 13 times when sentences were coded as error-free by the human annotators. However, after looking at these 13 false positives, the authors have found that seven (54%) of them were not actually false positives, but rather true quantifier errors which were not noticed by annotators.

Of the six remaining false alarms, two (15%) were caused by Stanford CoreNLP assigning the wrong POS tag to the sentences. In these cases,

CyWrite either failed to capture learner errors, or captured them incorrectly. For example, the sentence 'Also grand parents can raise their babies for couples who both work.' did not contain a quantifier error. However, the Stanford CoreNLP tagged the word 'work' as a noun. A construction that included 'both' followed by a singular noun alone without the 'and' conjunction with another noun in it was treated as a quantifier error by CyWrite. In this example, the incorrect POS tag caused a false positive. Fortunately, the frequency of such false positives was very low, because Stanford CoreNLP in general appears to be extremely robust.

Four (30%) false positives were due to problems with our hand-written rules. The rules rely on lists of uncountable nouns and nouns that could be used both as countable and uncountable. If one of the lists lacks a word that should in fact be there, this may cause a false positive. For example, the sentence 'So I will choose a career which I have much interest' was correct in terms of quantifier use. However, the noun 'interest' could be used both as a countable and uncountable noun, but we failed to include it in the relevant list for CyWrite, which caused a false positive. While such false positives hurt the performance of CyWrite, most of them they are easily fixable by either revising the rules or updating the word lists that accompany the rules.

Regarding the false negatives, our findings indicate that CyWrite failed to capture a total of 22 quantifier errors flagged by the annotators. As a result of a deeper analysis, we found that false negatives were caused by rule issues (N = 9, 41%), Stanford CoreNLP failures (N = 2, 9%), and learner language problems (N = 1, 5%). A portion of false negatives were actually human failures (N = 10, 45%) as they represent false alarms raised by the annotators, while CyWrite made correct determination in these cases.

Issues with subject-verb agreement error detection

CyWrite's detection of subject-verb agreement errors yielded 53 false positives and 40 false negatives. Regarding the false positives, CyWrite found 53 instances which were actually judged as correct usage of subject-verb agreement. An in-depth analysis showed that these false positives were caused by learner language problems (N = 18, 34%), annotation mistakes (N = 16, 30%), parser failures (N = 12, 23%), and rule issues (N = 7, 13%).

One third of false positives were in fact annotation mistakes. Learner language problems, such as spelling errors and ill-formed verbs, also influenced the accuracy of the detection. For example, a learner sentence, 'By the way, this process shuld be through the whole preparation period' was detected as an error, because 'should' was misspelled as 'shuld'; therefore, CyWrite detected 'shuld' as a singular noun, which cannot precede a copula *be* directly. Additionally, learner sentences that contain ill-formed verbs would also cause

incorrect detection. In this learner sentence 'In Parks library, computers are exists each floor for searching information', CyWrite detected 'computers' as the subject and 'exists' as the verb. However, due to the ill-formed verbs, this is an inaccurate detection.

Furthermore, parser failures led to 25% of inaccurate detection. As in this sentence, 'Sometimes, there are many snakes and mouse in their house or office', 'are' was used correctly. However, because the parser tagged 'mouse' as a singular noun, this sentence was detected as an error. In another example, 'Also, there is a concept that is "Self-service Grab-and-go", where you can make yourself sandwiches and wraps, assorted side and entree salads, and dessert, on the other hand, it guarantees high efficiency', because the parser recognized 'wraps' as a third-person singular verb, it sees this instance as an error, but it was in fact not a subject-verb agreement error.

Lastly, one sentence structure was not excluded in the rules, which caused some false positives as well. CyWrite already excludes the case where causative verbs (get, help, let, and make) can precede base form of verbs (e.g., ' … and make other people follow his mind'). Nevertheless, the false positives happen when there was an adjective or a past participle instead of a noun after the second verb. Sentences such as, 'Different cultures make this city become more diversified and colorful', and 'He had to give up his most free time to make the study not be influenced', have a slightly different structure from the previous one. Fortunately, this rule can be added to CyWrite easily to eliminate these false positives.

As for the false negatives, our results showed that CyWrite failed to capture 40 subject-verb agreement errors which were detected by annotators. From the in-depth analysis, we found similar patterns of false negatives: rule issues ($N = 20$, 50%), parser failures ($N = 9$, 23%), annotation mistakes ($N = 6$, 15%), and learner language problems ($N = 5$, 13%). Half of the false negatives were caused by rule issues. In order to achieve higher precision, certain structures were excluded from the rules, which impacted recall.

Issues with article error detection

With respect to error detections by CyWrite, we found 182 false positives and 72 false negatives. In the case-by-case analysis, we found that 26 (14%) 'false positives' were in fact annotator errors; others were due to issues of article error detection rules ($N = 104$, 76%), incorrect POS tagging ($N = 29$, 16%), errors caused by learner language issues, such as spelling ($N = 14$), and some other errors ($N = 9$) that we could not classify due to the ambiguity of the sentences. The high number of false positives caused by rule issues was due to two reasons: (1) our list of uncountable nouns was incomplete; and (2) rules did not account for all relevant syntactic structures. For instance, noun phrases

with 'what' as in 'Thus, what situation you are under is very important' are determined with question words and are correct in terms of article use. However, because this specific structure had not been defined in the article error detection rules, CyWrite detected such phrases as an error. Such false positives are very important and informative for improving the rules and increasing the precision and recall of CyWrite.

Regarding the article errors that CyWrite failed to capture, the findings yielded 72 false negatives. Similar to the explanations above, while 35 (49%) of these false negatives were due to annotation errors, 31 (43%) were because of the rule issues. Three (4%) errors were related to incorrect parser tagging and we were unable to categorize three (4%) other errors because of ambiguity.

Issues with run-on sentence detection

CyWrite detected run-on sentence errors and generated 19 false positives and 71 false negatives. Regarding the false positives, CyWrite found 19 instances which were actually error-free of run-on sentences. An extensive analysis showed that these false positives were caused by learner language problems (N = 9, 47%), annotation mistakes (N = 8, 42%), parser failures (N = 1, 5%), and rule issues (N = 1, 5%).

Half of false positives were produced because of learner language problems for missing a space and missing punctuations. For example, in this sentence, 'It is really helpful for students to search some information. Of course, the library provides books for students to study as well', a space is omitted between the period '.' and 'Of'; in this case, the parser could not parse the sentence correctly which led to an incorrect detection by CvWrite. Also, the sentence 'She says, "I went to attend meeting for giving lecture and sharing some experience with oral health professionals' lacks a double-quotation ' " ' mark at the end of the sentence, which generated the false positive.

The other half of 'false positives' was in fact annotation mistakes. In other cases, the parser failed in that it did not recognize the use of the transitional phrase 'that is to say'; and there was a rule issue because the structure of using the 'to-infinitive' in the beginning of a sentence was not included as a rule.

As for the 71 false negatives that were not captured by CyWrite, the causes were rule issues (N = 40, 56%), parser failures (N = 14, 20%), learner language problems (N = 10, 14%), and annotation mistakes (N = 7, 10%). Similar to the rules for subject-verb agreement errors, in order to maintain high precision, some sentence structures were suppressed because they were found to trigger more incorrect detections than accurate ones. More than half of the false negatives were caused by rule issues. Wrong POS tagging and wrong attachment of if-structure by the parser caused 20% of false negatives. These cases will be further analyzed and a solution will be sought.

3.3. Summary

In this section, we presented the precisions, recalls, F-scores, and accuracy of CyWrite and Criterion, and the reliability between human annotators for each error, and we also provided an in-depth analysis of the sources of false positives and negatives in order to further enhance CyWrite rules. Generally, CyWrite error detection on quantifiers, subject-verb agreement, articles, and run-on sentences performed better than Criterion in terms of the F-scores. While quantifier detection reached a higher precision than Criterion, the other three error detection achieved higher recalls. From the analysis of the sources of CyWrite failures, most of the false positives were from rule issues and annotation mistakes, whereas most of the false negatives were from annotator mistakes and learner language problems. This information will inform the improvement of CyWrite rules.

4. Conclusions

In this paper, we aimed to answer two research questions: (1) How well does CyWrite detect grammatical errors in ESL students' academic writing? and (2) If CyWrite's performance is not perfect, what are the reasons for that and how the performance can be improved? In this design-based study, we annotated a large sample of learner texts to serve as the gold standard for error detection, and revealed the possibility of developing software for grammatical error detection using a hybrid (statistical and rule-based) NLP approach. Echoing Leacock *et al*. (2010), our findings show that using a hybrid system combining a statistical approach (i.e., parsing of sentences) and rule-based approach (hand-coded linguistic rules) enables grammatical error detection to perform better than the sole use of a statistical approach. This study also reported accuracy, precision, recall, and F-scores; used more than one annotator and reported the reliability between human annotators; described the nature of the corpus; illustrated the targets of the intended grammar errors; and provided inclusion and exclusion of the error features. Of particular note is that the performance of CyWrite detection on the four target error types, quantifiers, subject-verb agreement, articles, and run-on sentences, outperformed Criterion's. This might be partially due to mismatched definitions of errors employed by the two systems. In addition, although the development and the testing corpora were kept strictly separate throughout our work, they were sampled from a bank of comparable writing, which might have also given CyWrite an edge in the comparison to Criterion. From the DBR perspective, it is more important that the analysis of CyWrite failures demonstrates that that achieving better performance is possible once the issues discovered in this paper are addressed.

4.1. CyWrite software revision plan

This paper presents the initial developmental stage of DBR in automated grammatical error detection. Therefore, it is necessary to apply what we have learned from our results to generate a revision plan for the next iteration. From the analysis of false positives and false negatives, we are able to make concrete plans for our refinement of the manually developed rules. First of all, we conclude that our quantifier rules were sufficiently accurate in the first iteration, but our findings can be still used for further enhancement. For the subject-verb agreement and run-on sentences rules, we learned that it is difficult to strike the balance between achieving higher precision and higher recall. Lastly, the relevant wordlists for article rules should be expanded, and the grammatical structures identified by the present study should be worked into our ruleset.

Another aspect that we need to focus on in the next iteration is the annotation mistakes. The quality of annotation in this study was acceptable in terms of α-reliability: $\alpha > 71$ which is considered 'good' reliability (Strijbos & Stahl, 2007). Nevertheless, the incorrect annotations introduced noise into the performance metrics and slowed down the development process. We conclude that a higher standard should be set for the reliability of human annotation in this kind of application. The incorrect annotations in our study could have been caused by insufficient calibration and annotators' fatigue; thus, we will train our future annotators with more guidance and a higher standard, and assign a fixed number of sentences for a certain amount of time when they conduct annotations to avoid overstrain. In addition, we may reconsider the decision to use a random subsample of the data for reliability purposes, and instead revert to having at least two annotators assigned to each sentence in the corpus. This way, annotators could be required to reach an agreement on every discrepancy, thus ruling out some of the false positives and false negatives that might be due to human error.

4.2. Pedagogical implications

AWE systems have been developed to assist language-learning students' writing by mainly providing grammatical feedback. To keep students' trust in AWE tools and to lower students' cognitive loads, high precision of rule detection is being maintained (Yuan & Felice, 2013). In our study, analyzing the sources of false positives and negatives helps improve rules to achieve this goal. Only when the detection is accurate, it is then possible to provide accurate and suitable feedback for students to revise their writing.

Additionally, the rule-based approach can facilitate not only grammatical error detection, but also causal discourse analysis (Chukharev-Hudilainen & Saricaoglu, 2014). Therefore, besides detecting language learners' grammatical

errors, detecting discourse-specific features (e.g., phraseology and verb tenses used in research articles) could provide students with additional contextual feedback.

4.3. Future research

As more types of grammatical errors are being detected by CyWrite, we plan to conduct improved versions of the present research study to evaluate CyWrite's performance in the detection of these errors. Following each iteration, the accuracy of the system can be improved, which could benefit L2 writing in the long run. However, as Leacock *et al.* (2010) pointed out, even though the accuracy of AWE tools can be evaluated and improved, the question of whether the systems assist language learners to improve their writing skills may remain unanswered for some time. Therefore, in addition to the accuracy of the grammatical error detection, research on automated feedback provision, student uptake of feedback, and the effect thereof on students' writing performance should also be carried out to maximize the benefits for ESL learners from any AWE tool.

Acknowledgements

We would like to acknowledge the contribution of all CyWrite project team members to corpus annotation, and especially Miles J. Conlan who worked on the troubleshooting and enhancement of the run-on sentence detection rules. We are also grateful to Joe Geluso and the two anonymous reviewers for their helpful comments on the early drafts of this manuscript.

Note

1. Because Criterion detects certain types of article errors not detected by CyWrite, sentences flagged as false-positive detections were re-annotated so that Criterion's precision on this error type would not be underestimated.

About the authors

Hui-Hsien Feng is a postdoctoral research associate at Iowa State University. She holds an MA in TESOL at the Ohio State University and a PhD in Applied Linguistics and Technology at Iowa State University. Her research interests include computer-assisted language learning, second-language academic writing, automated writing evaluation, and computational linguistics. She has presented her work at local and international conferences, including Second Language Research Forum (SLRF), the Computer-Assisted Language Instruction Consortium (CALICO), Technology for Second Language Learning (TSLL), Symposium on Second Language Writing (SSLW), and MIDTESOL. hhfeng@iastate.edu, Iowa State University.

Aysel Saricaoglu is an assistant professor in the English Language Education program at TED University (Ankara, Turkey). She holds a BA in English Language Teaching from Mersin University, Turkey and an MA in Teaching English as a Foreign Language from Hacettepe University, Turkey. After she worked as a foreign language instructor for three years, she was awarded the Fulbright scholarship, and she pursued her PhD degree in Applied Linguistics and Technology at Iowa State University. She taught academic writing to undergraduate and graduate non-native speakers of English and also worked in the CyWrite (a web-based automated writing evaluation tool) project. Her research interests include computer-assisted language learning, automated writing evaluation, computational linguistics, academic writing, and systemic functional linguistics. saricaogluaysel@gmail.com, TED University.

Evgeny Chukharev-Hudilainen is an assistant professor in the Applied Linguistics and Technology program at Iowa State University. He holds BSc and MSc degrees in computer science and engineering from Northern Federal University of Russia, and a PhD in applied and computational linguistics from Herzen State Pedagogical University. Prior to joining ISU in 2012, Evgeny spent more than six years working as a senior software engineer at the Central Bank of Russia. His current research interests lie at the intersection of applied linguistics, psycholinguistics, and computational linguistics. He is leading the development of CyWrite, a web-based automated writing evaluation tool, and teaching graduate and undergraduate courses in linguistics at ISU. evgeny@iastate.edu, Iowa State University.

References

Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research?. *Educational Researcher*, 41 (1), 16–25. http://dx.doi.org/10.3102/0013189X11428813

Attali, Y. (2004). *Exploring the feedback and revision features of Criterion*. Paper presented at the National Council on Measurement in Education Annual Meeting, San Diego, CA.

Bender, E. M., Flickinger, D., Oepen, S., Walsh, A. & Baldwin, T. (2004). ARBORETUM: Using a Precision Grammar for Grammar Checking in CALL. In *Proc. InSTIL/ICALL Symposium on Computer Assisted Learning*, Venice, Italy.

Bitchener, J., & Ferris, D. (2012). *Written corrective feedback in second language acquisition and writing*. New York: Routledge.

Bramer, M. (2013). *Logic programming with prolog*. London: Springer. http://dx.doi.org/10.1007/978-1-4471-5487-7

Burstein, J. (2003). The e-rater scoring engine: Automated Essay Scoring with natural language processing. In M. D. Shermis and J. C. Burstein (Eds), *Automated Essay Scoring: A cross disciplinary approach,* 113–121. Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J., Chodorow, M., & Leacock, C. (2004). *Criterion*^{SM} online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the Fif-*

teenth Annual Conference on Innovative Applications of Artificial Intelligence, Acapulco, Mexico. Retrieved from http://www.ets.org/research/policy_research_reports/publications/chapter/2004/cwjd

Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The grammar book: An ESL/EFL teacher's course*. Boston, MA: Heinle & Heinle.

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, 32 (2), 385–405.

Chukharev-Hudilainen, E., & Saricaoglu, A. (2014). Causal discourse analyzer: Improving automated feedback on academic ESL writing. *Computer Assisted Language Learning*. http://dx.doi.org/10.1080/09588221.2014.991795

Connors, R. J., & Lunsford, A. A. (1988). Frequency of formal errors in current college writing, or Ma and Pa Kettle do research. *College Composition and Communication*, 395–409. http://dx.doi.org/10.2307/357695

Cowan, R. (2008). *The teacher's grammar of English: A course book and reference guide*. Cambridge: Cambridge University Press.

Craig, J. L. (2013). *Integrating writing strategies in EFL/ESL university contexts: A writing-across-the-curriculum approach*. New York: Routledge.

Davies, M. (2010). The corpus of contemporary American English as the first reliable monitor corpus of English. *Language and Literary Computing,* 25 (4), 447–464. http://dx.doi.org/10.1093/llc/fqq018

De Feliece, R. (2008). *Automatic error detection in non-native English* (Unpublished doctoral dissertation).University of Oxford, England.

De Marneffe, M.C., MacCartney, B., & Manning, C.D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC* (Vol. 6), 449–454. Genoa: ELRA.

DeCapua, A. (2008). *Grammar for teachers: A guide to American English for native and non-native speakers.* Boston, MA: Springer. http://dx.doi.org/10.1007/978-0-387-76332-3

Echevarria, J., Short, D., & Powers, K. (2006). School reform and standards based education: A model for English-language learners. *The Journal of Educational Research*, 99, 195–210. http://dx.doi.org/10.3200/JOER.99.4.195-211

Ferris, D. R. (1999). The case for grammar correction in L2 writing classes. A response to Truscott (1996). *Journal of Second Language Writing,* 8, 1–10. http://dx.doi.org/10.1016/S1060-3743(99)80110-6

Ferris, D. R. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In K. Hyland & F. Hyland (Eds), *Feedback in second language writing: Contexts and issues,* 81–104. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139524742.007

Ferris, D. R. (2010). Second language writing research and written corrective feedback in SLA: Intersections and practical applications. *Studies in Second Language Acquisition,* 32, 181–201. http://dx.doi.org/10.1017/S0272263109990490

Ferris, D., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing,* 10, 161–184.

Fitzpatrick, M. (2011). *Engaging writing 2: Essential skills for academic writing*. White Plains, NY: Pearson Education.

Guide to Grammar and Writing. (n.d.). Retrieved from http://grammar.ccc.commnet.edu/grammar/

Hagerman, C. (2011). An evaluation of Automated Writing Assessment. *JALT CALL Journal,* 7 (3), 271–292.

Hayes, A.F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1 (1), 77–89. http://dx.doi.org/10.1080/19312450709336664

Heift, T. (2004). Corrective feedback and learner uptake in CALL. *ReCALL,* 16 (2), 416–431. http://dx.doi.org/10.1017/S0958344004001120

Hinkel, E. (2011). What research on second language writing tells us and what it doesn't. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning, Volume 2,* 523–538. New York: Routledge.

Hung, H-T. (2011). Design-based research: Designing a multimedia environment to support language learning. *Innovations in Education and Teaching International*, 48, 159–169. http://dx.doi.org/10.1080/14703297.2011.564011

Kelly, A. E., Lesh, R. A., & Baek, J. Y. (2008). *Handbook of design research methods in education: Innovations in science, technology, engineering, and mathematics learning and teaching*. New York: Routledge.

Kennedy-Clark, S. (2013). Research by design: Design-based research and the higher degree research student. *Journal of Learning Design*, 6 (2), 26–32. http://dx.doi.org/10.5204/jld.v6i2.128

Klein, D., & Manning, C.D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics – Volume 1*, 423–430. Association for Computational Linguistics. Morristown, NJ: ACL.

Krippendorff, K. (2011). Computing Krippendorff's alpha reliability. *Departmental Papers (ASC)*, 43, 1–10.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Automated grammatical error detection for language learners. *Synthesis Lectures on Human Language Technologies,* 3 (1), 1–134. http://dx.doi.org/10.2200/S00275ED1V01Y201006HLT009

Li, J., Link, S, & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation in ESL writing instruction. *Journal of Second Language Writing*, 27, 1–18. http://dx.doi.org/10.1016/j.jslw.2014.10.004

Li, Z., Feng, H.-H., & Saricaoglu, A. (in press). *The short-term and long-term effects of AWE feedback on ESL learners' development of grammatical accuracy. CALICO Journal.*

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System*, 44, 66–78. http://dx.doi.org/10.1016/j.system.2014.02.007

Link, S., Dursun, A., Karakaya, K., & Hegelheimer, V. (2014). Towards better ESL practices for implementing automated writing evaluation. *CALICO Journal,* 31 (3). http://dx.doi.org/10.11139/cj.31.3.323-344

Lund, A. (2005). Collective epistemologies in an upper secondary school. A preliminary analysis. Paper presented at the *EARLI conference,* Nicosia, CY.

Lund, A. (2008). Wikis: A collective approach to language production. *ReCALL*, 20 (1), 35–54. http://dx.doi.org/10.1017/S0958344008000414

Lund, A. & Smordal, O. (2006) Is there a space for the teacher in a Wiki? In: *Proceedings of the 2006 International Symposium on Wikis* (WikiSym '06), 37–46. Odense, Denmark: ACM Press.

Lunsford, A. A. (2012). *The everyday writer* (5th ed.). Boston, MA: Bedford/St. Martins.

Lunsford, A. A., & Lunsford, K. J. (2008). 'Mistakes are a fact of life': A national comparative study. *College Composition and Communication,* 59 (4), 781–806.

Meurers, D. (2012). Natural language processing and language learning. In C. A. Chapelle, (Ed.), *Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell.

Nassaji, H., & Fotos, S. (2011). *Teaching grammar in second language classrooms. Integrating form-focused instruction in communicative context*. London: Routledge.

O'Donnell, M. (2008). The UAM CorpusTool: Software for corpus annotation and exploration. In *Proceedings of the XXVI Congreso de AESLA*, Almeria, Spain.

Pardo-Ballester, C., & Rodríguez, J. C. (2009). Using design-based research to guide the development of online instructional materials. In C. A. Chapelle, H. G. Jun, & I. Katz (Eds), *Developing and evaluating language learning materials,* 86–102. Ames, IA: Iowa State University.

Pardo-Ballester, C., & Rodríguez, J. C. (2010). Developing Spanish online readings using design-based research. *CALICO Journal*, 27, 540–553. http://dx.doi.org/10.11139/cj.27.3.540-553

Pardo-Ballester, C., & Rodríguez, J. C. (2013). Design principles for language learning activities in synthetic environments. In J. Rodríguez & M. Pardo-Ballester (Eds), *Design-Based Research in CALL,* 183–209. San Marcos, TX: CALICO.

Plomp, T. (2009). Educational design research: An introduction. In T. Plomp & N. Nieveen (Eds), *An introduction to educational design research,* 9–35. Enschede: The Netherlands: SLO Netherlands Institute for Curriculum Development.

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2014). *AWE for formative assessment: Investigating accuracy and efficiency as part of argument-based validation*. Paper presented at The 3rd Teachers College, Columbia University Roundtable in Second Language Studies, New York.

Richards, J. C., & Rodgers, T. S. (2001). *Approaches and methods in language teaching* (2nd ed.). Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511667305

Rodríguez, J. C. & Pardo-Ballester, C., (2013) (Eds). *Design-based Research in CALL*. CALICO Monograph Series, Volume 8. San Marcos, TX: CALICO.

Run-on sentence. (n.d.). In *Merriam-Webster Online Dictionary*. Retrieved from http://www.merriam-webster.com/dictionary/run-on+sentence

Schneider, D., & McCoy, K. F. (1998). Recognizing syntactic errors in the writing of second language learners. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2,* 1198–1204. Association for Computational Linguistics. http://dx.doi.org/10.3115/980432.980765

Shutler, R. (2012). *A study of student and teacher perceptions of criterion, an online writing program* (Unpublished master's thesis). Carleton University: Ottawa, CA.

Strijbos, J.-W., & Stahl, G. (2007). Methodological issues in developing a multi-dimensional codingprocedure for small-group chat communication. *Learning and Instruction*, 17 (4), 394–404. http://dx.doi.org/10.1016/j.learninstruc.2007.03.005

Sun, X. (2014). Analysis on negative transfer of native language syntax structure in English compositions of Chinese college students. In *3rd International Conference on Science and Social Research (ICSSR 2014)*. http://dx.doi.org/10.2991/icssr-14.2014.229

The Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher*, 32 (1), 5–8. http://dx.doi.org/10.3102/0013189X032001005

Wang, F., & Hannafin, M. J. (2005). Design-based research and technology enhanced learning environments. *Educational Technology Research and Development,* 53 (4), 5–23. http://dx.doi.org/10.1007/BF02504682

Wu, H. P., & Garza, E. V. (2014). Types and attributes of English writing errors in the EFL context – A study of error analysis. *Journal of Language Teaching and Research*, 5 (6), 1256–1262. http://dx.doi.org/10.4304/jltr.5.6.1256-1262

Yuan, Z., & Felice, M. (2013). Constrained grammatical error correction using statistical machine translation. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning (CoNLL 2013): Shared Task*, 52–61. Madison, WI: Omnipress.

Yutdhana, S. (2005a). Design-based research in CALL. In J. L. Egbert & G. Mikel Petrie (Eds), *CALL research perspectives,* 169–178. Mahwah, NJ: Lawrence Erlbaum Associates.

Yutdhana, S. (2005b). *The development of a teacher-training for model in using the Internet for teaching English as a foreign language*. (Unpublished doctoral dissertation). Suranaree University of Technology, Thailand.