LANGUAGE
TESTING

Language Testing
27(3) 419–436

DOI: 10.1177/0265532210364391
http://ltj.sagepub.com
SAGE

The utility of article and preposition error correction systems for English language learners: Feedback and assessment

Martin Chodorow
Hunter College of the City University of New York, USA

Michael Gamon
Microsoft Research, USA

Joel Tetreault
Educational Testing Service, USA

Abstract

In this paper, we describe and evaluate two state-of-the-art systems for identifying and correcting writing errors involving English articles and prepositions. *Criterion*[SM], developed by Educational Testing Service, and *ESL Assistant*, developed by Microsoft Research, both use machine learning techniques to build models of article and preposition usage which enable them to identify errors and suggest corrections to the writer. We evaluated the effects of these systems on users in two studies. In one, *Criterion* provided feedback about article errors to native and non-native speakers who were writing an essay for a college-level psychology course. The results showed a significant reduction in the number of article errors in the final essays of the non-native speakers. In the second study, *ESL Assistant* was used by non-native speakers who were composing email messages. The results indicated that users were selective in their choices among the system's suggested corrections and that, as a result, they were able to increase the proportion of valid corrections by making effective use of feedback.

Keywords

computer assisted language learning, ESL writing errors, grammatical error detection, natural language processing, proofing tools

The National Clearinghouse for English Language Acquisition estimates that 9.6% of the students in the US public school population speak two or more languages and have limited

Corresponding author:
Martin Chodorow, Hunter College of the City University of New York, New York, NY 10065, USA.
Email: martin.chodorow@hunter.cuny.edu

English proficiency. In most states, this number is growing. For example, in New York, it increased by 22% from 1990 to 2000 (http://www.ncela.gwu.edu/policy/states/reports/statedata/2001). At the university level, there are estimated to be more than half a million international students in US colleges and universities whose native language is not English (Burghardt, 2002). In fact, non-native English writers are a large and growing segment of the world's population. Estimates are that in China alone as many as 300 million people are currently studying English. Clearly, there is a substantial and increasing need for tools to support English language learning and instruction at all levels and in all countries.

Among the most difficult elements of English for the non-native speaker to master are articles (*a/an*, *the*, or no article) and prepositions (*of*, *to*, *with*, *at*, etc.). Together, it is claimed that they account for 20%–50% of all grammar and usage errors made by learners of English as a Second Language (ESL) (Dalgish, 1985; Diab, 1997; Izumi et al., 2003; Bitchener et al., 2005). Articles and prepositions are so difficult, in part because their usage depends on the interaction of many heterogeneous factors. For example, selection of an article may depend on the noun that follows it and on the words that modify that noun. Compare the phrase *'a damage', which is unacceptable, to the phrase 'a little damage', which is fine. Discourse factors often affect the choice of *a* versus *the*; first mention of an entity is usually marked with *a* and subsequent references to it are marked with *the*. For prepositions, the selection can depend on the noun which follows the preposition ('in the tank'), the verb which precedes it ('drained off'), the noun which precedes it ('a gallon of'), or a combination of all three ('I drained a gallon from the tank').

Articles and prepositions are also difficult to learn because of variability in their usage. Judging the correctness of article and preposition usage is often hard even for native English speakers. In a recent study (Tetreault and Chodorow, 2008b), we found the judgments of two trained native speakers to differ by as much as 10% when rating preposition usage as correct or incorrect in essays written by non-native speakers on the Test of English as a Foreign Language (TOEFL®). This variability stands in stark contrast to the consistency of native speakers' judgments about most grammatical errors, which are typically clear-cut violations of a rule of syntax, such as the requirement that the subject and the verb must agree in number (*'The boys is here'). In Tetreault and Chodorow (2008b), we also reported that when two native English speakers were asked to fill in the best preposition in a collection of 200 sentences with one preposition 'blanked out', there was 24% disagreement. In summary, the complexity and variability of article and preposition usage make these elements of English syntax difficult for human learners and even more difficult for automated error-detection and correction systems.

In two separate projects (Han et al., 2006; Chodorow et al., 2007; Tetreault and Chodorow, 2008a; Gamon et al., 2008; Gamon et al., 2009), we have developed automatic methods for detecting and correcting article and preposition errors as part of larger systems designed to provide feedback to writers about the quality of their grammar, word usage, and mechanics (spelling, punctuation, etc.). The first system has been developed by Educational Testing Service (Burstein et al., 2004) as part of the *Criterion*℠ online writing evaluation service. We refer to this system as *Criterion*. Article error detection was added to *Criterion* in 2005, and preposition error detection was added in 2008. The second system is the Microsoft Research *ESL Assistant* (Gamon et al., 2009), which was made public as a prototype web service in June of 2008. Both

Criterion and *ESL Assistant* provide their users with a full range of grammatical error detection and correction – although only their article and preposition components are described here. *Criterion* and *ESL Assistant* primarily target the usage of Standard American English, but it should be noted that their design as data-driven systems makes them highly configurable to different English dialects, as long as sufficient amounts of training data are available.

Evaluation of a system for detecting and correcting usage errors can be performed along at least two dimensions: system-centric evaluation, and user-centric evaluation. In terms of Chapelle's (2001) criteria for evaluating Computer Assisted Language Learning (CALL) applications, these correspond to the practicality of an application and its impact, respectively. The former focuses on the performance of the system itself, for example, on statistics about the number of correctly identified errors. In contrast, the latter examines the user's behavior or the effect of the system on the user. System-centric results for the two systems have been reported in detail in our earlier studies. In this paper, we present the methods we have used to build these large, state-of-the-art systems and evaluate the effects of providing error feedback to the writer. In particular, we focus on user-centric evaluation and ask how our systems impact the quality of the user's writing. In one study, we examine the benefits of error feedback in writing an essay for a college-level psychology course, and, in another, we look at the value of error feedback in writing email. To our knowledge, there have been no prior user-centric evaluations of systems that detect article and preposition errors.

Summary of Approach

Criterion and *ESL Assistant* view the process of detecting an error in article or preposition usage as a classification task. A context in which an article or preposition may appear is analyzed as a set of features, such as a sequence of words and parts of speech (PoS). Based on the features, the context is assigned to one or more classes. For example, the context 'We sat ___ the sunshine', might be assigned to the preposition class 'in' but not to the class 'at'. If this context occurs with 'at' filling the blank, then the classifier marks it as an error. Some error classifiers have been assembled from sets of handcrafted rules (Eeg-Oloffson and Knuttson, 2003), but building a classifier by hand has proven to be both time-consuming and costly. By contrast, corpus-based statistical approaches to classification (Han et al., 2006; Nagata et al., 2006; de Felice et al., 2007; Turner and Charniak, 2007) are less labor intensive and permit the construction of classifiers that can cover a wider range of contexts and adapt more readily to differences in dialects, topical content and genres of writing.

In place of handcrafted rules, statistical classifiers use measures of statistical association between features and classes based on frequencies of occurrence in large corpora of published text, such as collections of newspapers, reference works, and textbooks. One of the best ways for a person to learn preposition and article usage is through continued exposure to many examples of correct usage, and statistical classifiers are trained on millions of examples of correct usage. Of course, when people are formally taught a second or foreign language, they are also shown many examples of errors to avoid. We would like to train our classifiers on both correct usage and on errors, but, unfortunately, most

of the available corpora of learners' errors are too small for the purpose of statistical training. A related problem is that learners' writing is often riddled with all sorts of grammar, usage, and style errors that make it difficult for humans and automatic systems alike to determine just what error a given sentence illustrates. For example, does the string 'I fond car' contain a misspelling of 'found' and an omitted article ('I found a car'), or is it lacking a main verb, the preposition 'of', and a marker for plural ('I am fond of cars')? Because of the scarcity of learner corpora and the ambiguity of learners' errors, corpus-based statistical approaches have thus far restricted their training to cases of correct usage.

Contextual representations: Transforming text into a set of features

Both systems use the same general approach to representing a context of usage as a set of features. Although *ESL Assistant* and *Criterion* have many similarities, they were developed independently and generate their contextual representations using different tools. In the examples and discussion below, we will focus on prepositions, but the same techniques are used for articles as well.

To train a classifier, each context of usage in the training corpus must be represented as a set of features. As a first step in constructing these features (see Example 1), the words of the text are automatically tagged with their PoS (Ratnaparkhi, 1997), such as pronoun (PRP), past tense verb (VBD), article (DT), preposition (IN), and singular noun (NN).The tagged words are then organized into larger units ('chunks'), such as noun phrases (NP) and verb phrases (VP), shown in square brackets in example (1).

(1) [NP We_PRP] [VP sat_VBD [PP at_IN [NP the_DT sunshine_NN]]]

For each sentence in the training corpus, the contextual features are extracted from this tagged and bracketed representation as shown in Figure 1 and these features are fed into a classification training algorithm.

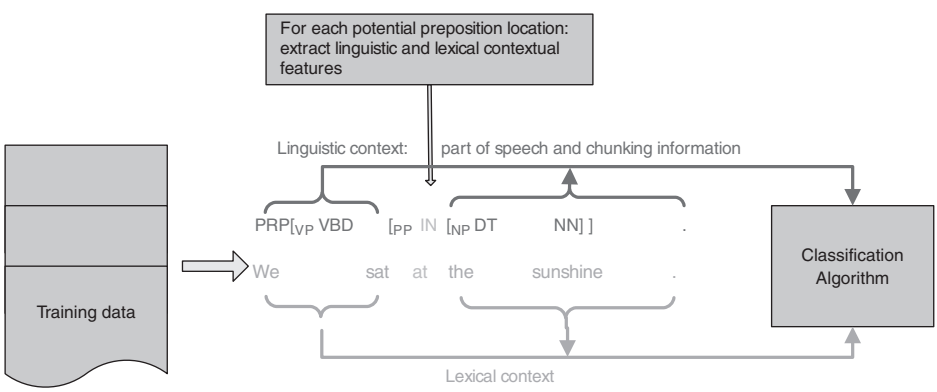

Figure 1. Contextual feature extraction and training for both systems

Table 1. Some features used in Criterion

Feature	Description	Example
TGLR	Preceding lemma with PoS tag and following lemma with PoS tag	sit_VBD−1+the_DT+1
TGL	Two preceding lemmas with PoS tags	we_PRP−2+sit_VBD−1
TGR	Two following lemmas with PoS tags	the_DT+1+sunshine_NN+2
BGL	Preceding lemma with PoS tag	sit_VBD−1
BGR	Following lemma with PoS tag	the_DT+1
FH	Lemma of headword of following phrase with PoS tag	sunshine_NN
FP	Following phrase including PoS tags	the_DT+sunshine_NN
FHword	Lemma of headword of following phrase	sunshine
PHR_pre	Preceding phrase type	VP
PV	Preceding verb lemma with PoS tag	sit_VBD
FHtag	PoS tag of headword of the following phrase	NN
PVtag	PoS tag of the preceding verb	VBD

The two systems differ in the specific set of features extracted from these representations. *Criterion* uses a set of 25 features to represent the context for prepositions, and 11 features for articles. Table 1 lists the name and description of some preposition features, as well as the value of the feature for example (1). All positions mentioned in the feature descriptions ('preceding', 'following', 'after') are relative to the position of the preposition. A 'lemma' is the base form of a word with the morphological inflections for number and tense removed (for example, 'sit' is the lemma of 'sat'). The features are listed in their order of importance to the statistical model, which means that the TGLR feature listed first contributes the most to the system's performance. *Criterion* is designed to handle a set of 34 prepositions.

In the *ESL Assistant*, a simple PoS-based heuristic determines whether a given position in the sentence is a potential location for a preposition. Potential locations for prepositions are defined as beginnings of noun phrases. For each such context, four tokens (words) to the left and to the right, and six PoS tags to the left and to the right are extracted as individual features, similar to work in contextual spelling correction (Golding et al., 1999). Each feature consists of a label (a PoS tag or a lexical item) followed by an indication of its position relative to the potential target position. 'PRP_-2', for example, indicates that two tokens to the left of the potential preposition position, there is a pronoun part-of-speech tag. In addition to this standard set of contextual features, a few 'custom' features that are designed to focus on salient properties of the context are added. The potential location of a preposition (in this case with the preposition *at* present in that position) in Figure 1 would generate the features in Table 2. The first four features are PoS features, and the remainder are contextual features.

The total number of features in the *ESL Assistant* is 28. Choice of prepositions is limited to a set of prepositions that are both frequent and figure prominently in non-native errors: *of*, *in*, *for*, *to*, *by*, *with*, *at*, *on*, *from*, *as*, *about*, *since*. Although there is a difference in the number of prepositions covered by the two systems, the extra prepositions that *Criterion* handles occur very infrequently when compared to the top 10 that both systems cover, which account for roughly 91% of preposition usage.

Table 2. Some features used in the *ESL Assistant*

Feature	Description	Example
PoSCL	Preceding PoS tags	PRP_−2,VP_−1
PoSCR	Following PoS tags	DT_+1, NN_+2, ._+3
LexCL	Preceding words and punctuation	we_-2, sat_-1
LexCR	Following words and punctuation	the_+1, sunshine_+2, ._+3
CapR/CapL	Presence of capitalized tokens to the left or right	<none available>
AcrR/AcrL	Presence of acronyms (tokens in all upper case)	<none available>
MN	Mass noun/count noun status of the headword of the NP	massNoun
HeadNP	Headword of the following NP	sunshine
HeadVP	Headword of the preceding VP	sat

Note that in both systems, while the number of features is relatively small, the number of feature *values* ranges in the hundreds of thousands. A simple contextual feature that uses the word which immediately precedes the preposition can assume a number of values that is roughly equal to the vocabulary size of the corpus. Consequently, both systems use techniques to reduce the number of feature values used in training.

Training a model of usage

To predict the most probable preposition given the feature values in a context, both systems use Maximum Entropy (ME) modeling (Ratnaparkhi, 1997), also known as multiple logistic regression. The ME model assigns weights to each of the features, and, when presented with a new context, the model uses the weights to compute a probability for each of the prepositions (or articles) it was trained on. The training data for both systems are listed in Table 3.

Constructing an end-to-end system

With the classifier constructed, we can develop a system to detect errors in learner writing. The system goes through several steps. First, the text is input into the system, and a *pre-processing step* PoS-tags, chunks and extracts features for each context. Next, each context is passed to the ME classifier which outputs a probability distribution over the set of prepositions.

Neither of the systems relies on the ME classifier alone. Since the goal of an error detection system is to provide diagnostic feedback, the system's outputs are heavily constrained to minimize false positives (i.e., the system tries to avoid saying the writer's preposition (or article) is used incorrectly when it is actually correct), and thus avoid misleading the writer.

In *Criterion*, after the ME classifier has output a probability for each of the 34 prepositions but before the system has made its final decision, a series of rule-based filters block what would otherwise be false positives. The *ESL Assistant* filters the set of potential corrections by adding a second source of statistical information. This filter is based

Table 3. Comparison of *Criterion* and *ESL Assistant*

	Criterion	*ESL Assistant*
Intended audience/use	Primarily high school and college students	Broad spectrum of users with web access
Deployment	Classroom Instructional Tool deployed in 2001; articles added in 2005; prepositions in 2008	Web-based application deployed in 2008
Feedback to the user	Explanations	Web-based examples
Targeted prepositions	34	12
Feature type	Contextual features	Contextual features
Feature vectors	25 features: 466,00 total feature values after feature reduction	28 features: part-of-speech and lexical features from a window of 6 and 4 tokens respectively around the preposition location 5 custom features 75,000 total feature values (after feature reduction)
Training data: sources	San Jose Mercury News, MetaMetrics Lexile corpus	Encarta, Reuters, UN corpus, Europarl corpus, web-scraped data
Training data: volume	3.1 million sentences	2.5 million sentences
Classifier	Maximum entropy	Maximum entropy
Post-processing	Rule-based filters, thresholds	Language Model scoring, thresholds

on a very large language model (i.e. a model that assigns probabilities to word sequences such as "He is teacher") that is trained on the Gigaword corpus (Linguistic Data Consortium, 2003), for details see Gao et al. (2001) and Nguyen et al. (2007). The language model provides a score for both the original user input and the potential correction. Only if the corrected version achieves a substantially higher language model score than the original user input is the suggested correction shown to the user.

Finally, both systems use various thresholds determined on a development corpus to refine the interaction between the ME classifier and the post-processing filters. These thresholds are set so as to strongly favor avoiding false positives even if it means reducing the overall number of errors that the system detects. A similar use of thresholds is found in Nagata et al. (2006) for article selection.

Table 3 summarizes the similarities and differences between the preposition error detection and correction components of *Criterion* and the *ESL Assistant*.

The role of web-based examples and meta-linguistic feedback

There is plenty of anecdotal evidence that non-native speakers use web search engines to find and verify usage of English expressions. The number of returned search results can serve as a proxy to identify 'correctness'. If, for example, a non-native speaker is confused about whether the common expression is 'on the other hand' or 'in the other hand', a quick string

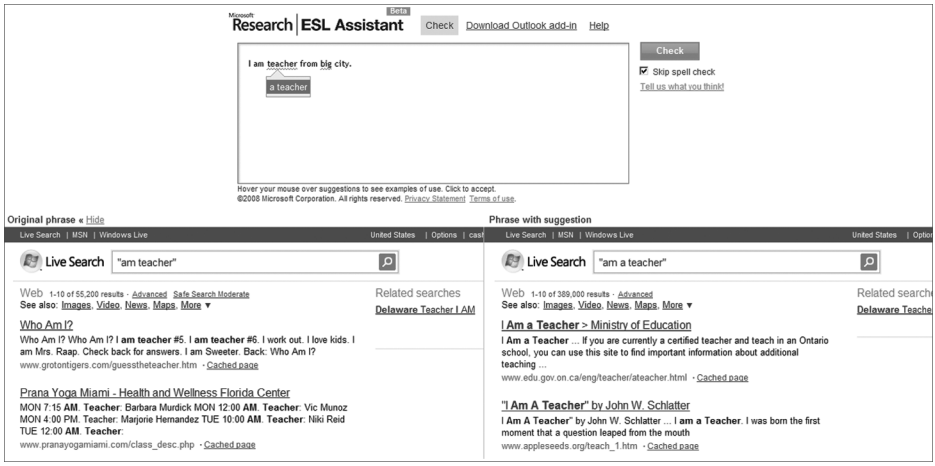

Figure 2. *ESL Assistant:* Side-by-side search results for original string and suggested correction

search using a major web search engine for both expressions will yield roughly 60 million hits for 'on the other hand' versus 250,000 hits for 'in the other hand'. Furthermore, inspection of the results will quickly confirm that 'on the other hand' is indeed the collocational expression, whereas 'in the other hand' – while being well-formed – is a literal expression.

The *ESL Assistant* incorporates this search for usage examples as illustrated in Figure 2. When a potential error is detected, and a suggested correction is offered, hovering over the suggested correction will trigger a side-by-side string search of both the original string and the suggested correction. This enables the user to verify whether the suggested correction corresponds to the intended usage. Note that this functionality is especially useful where the errors and potential corrections have a semantic component that will often leave the final usage decision to the user.

Figure 3 shows a screenshot of *Criterion* with a sample of its article error detection. *Criterion* provides some meta-linguistic feedback by labeling the type of error that the writer has made. It also makes available a more detailed explanation of the error type via the 'Writer's Handbook' link near the top of the screen, which will display the section of the online Handbook that is most relevant to the error. The value of meta-linguistic feedback has been documented by Heift (2004), who found that its inclusion resulted in fewer errors in students' re-submissions in a foreign language learning CALL task.

Evaluation

The two systems have been used in very different ways and for different purposes. *Criterion* is geared towards classroom use; the *ESL Assistant* is deployed as a prototype web application. Consequently, our evaluation will focus on different aspects of usage. The *Criterion* evaluation assesses whether the automatically provided corrections help students improve their writing over time, while the *ESL Assistant* evaluation will examine usage patterns logged from the web service.

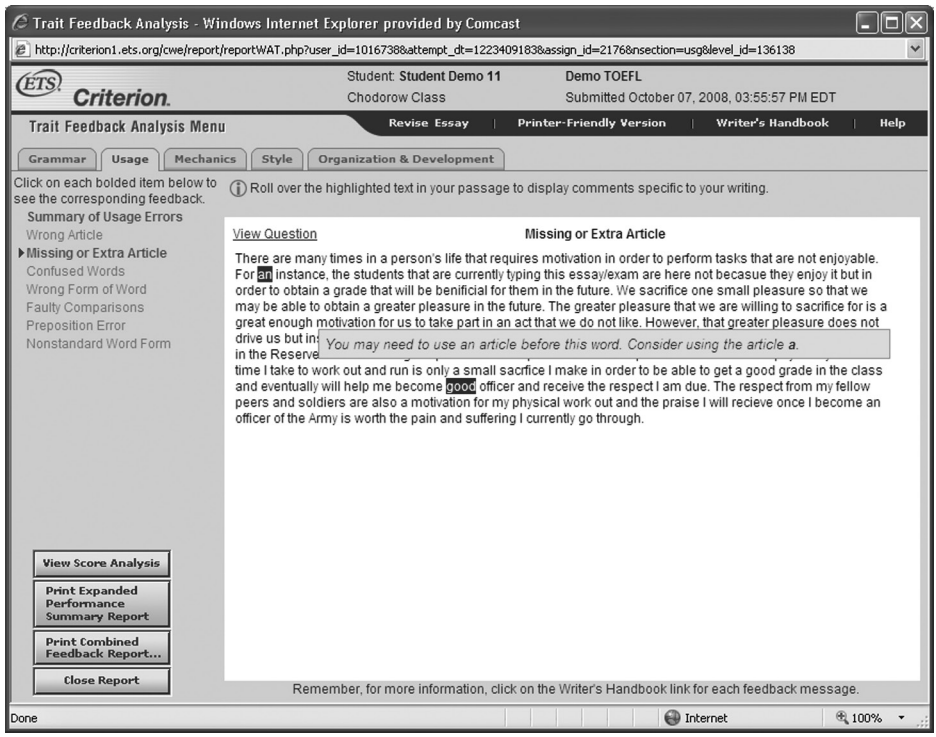

Figure 3. Screenshot of *Criterion* system

Experiment 1 (Criterion): Is feedback effective in improving the quality in student writing?

To date, most NLP research efforts have focused on the quality of error detection without considering the user. These evaluations typically involve a fixed collection of learner essays which are used to develop and evaluate a system. A human annotator marks the errors in the text, and the system's judgments are compared to the human's. A *hit* is a case that both the system and the human mark as an error. A *false positive* is a case that the system marks as an error but the human does not, and a *miss* is a case that the system does not mark as an error but the human does. Two measures are generally used to gauge system performance. *Precision* is the number of hits divided by the sum of the hits and false positives. *Recall* is the number of hits divided by the sum of the hits and misses. In detecting article errors, *Criterion*'s precision is about 90% and its recall is about 40%. That means that when the system reports an error in a student's writing, the human annotator agrees about 90% of the time. However, *Criterion* detects only about 40% of the errors that the human marks (Han et al., 2006). In detecting preposition errors, *Criterion*'s precision is about 80% and its recall is about 25% (Tetreault and Chodorow, 2008a). Once again, the disparity between precision and recall reflects our decision to maximize the former, even at the expense of the latter.

Arguably, system-centric measures, such as those given above, do not tell us the real value of an error detection system, which can only be measured in its effect on student writing over time. Optimally, one would want to chart a student's progress over a year or more while using a grammar error detection tool. Such longitudinal studies have not yet been conducted on *Criterion*, but at least one short-term study has: Attali (2004) analyzed 9000 first-draft and final-version essays written by American junior high and high school students using C*riterion.* The results indicated that, on average, errors were reduced in final submissions of essays for the error types that *Criterion* provides feedback on. However, the Attali study was conducted before article and preposition error detection had been incorporated into *Criterion*, and the overwhelming majority of the writers were native speakers of English.

In this section, we analyze essays from college students to measure the short-term effects of error detection on student writing. We compare the effects of feedback on native and non-native writers.

Participants. Our data come from a study by Lipnevich and Smith (2009), who examined the effects of feedback on students' examination performance. The participants were 463 undergraduate students enrolled in an introductory psychology course at two US universities.

Procedure. The participants were given the task of writing a 500-word essay on the topic of motivation, a content area in their course. This assignment was part of the course requirement, and the participants' grades on the essay counted toward their final grades for the semester. The participants wrote their essays in a computer laboratory. One week later, they reported back to the same computer laboratory for a second session. Some participants (N = 155) received no detailed feedback about their original essay ('No Feedback' condition), some (N = 159) were given detailed feedback that they were told had been generated by a computer ('Computer Feedback'), and some (N = 149) were provided detailed feedback that they were told had been generated by an instructor ('Simulated Feedback'). In both cases where feedback was given, the information about grammar, usage, mechanics, and style errors was actually produced by the *Criterion* writing system, which, at the time of data collection, contained a module for article error detection but not for proposition error detection. During the second session, all participants were encouraged to reread what they had written and to work on improving their essay. Following their revision of the essay, the second and final version of the essay was analyzed by *Criterion* and blindly graded by two human instructors. Lipnevich and Smith (2009) found that detailed feedback was strongly correlated with improvement in student essay scores.

Results and discussion. Of the 463 students in the study, 370 were native speakers of English and 93 were non-native speakers. For our analysis, we used the essays from students who made at least one article error in their original essay. This resulted in a set of essays from 268 native speakers and 71 non-native speakers. These non-native speakers had an article error rate of about 3%, which was lower than the 13% error rate reported by Han et al. (2006) for native speakers of Chinese, Japanese, and Russian who wrote essays for TOEFL. This difference is likely due to the fact that English article errors are most

common among native speakers of East Asian and Slavic languages, the groups studied by Han et al.; the non-native speakers in the Lipnevich and Smith (2009) study represented a wider range of languages. A second likely reason for the difference in rates is that all the students in the Lipnevich and Smith study had already achieved sufficient English proficiency to be admitted to a US university and were in an English immersion environment. By contrast, the TOEFL groups in Han et al. included many examinees who were studying English as a second or foreign language in a non-English-speaking country.

To determine if there were any differences between native and non-native speakers, and between students who received feedback and those who did not, we counted the number of article errors detected by *Criterion* in each student's original and revised versions of his/her essay. The data were then analyzed using an analysis of variance with essay version (original or revised) as a within-participants variable and language (native or non-native) and feedback (yes or no) as between-participants variables. (The 'yes' level of feedback consisted of the 'Computer Feedback' and 'Simulated Feedback' groups combined as there was no significant difference between them and no interactions with other variables.) There was a significant main effect for language, as non-native speakers had a higher overall error rate (M = 0.0333) than native speakers (M = 0.0285), $F(1,335) = 4.28$, $p = 0.039$, $\eta^2_{partial} = 0.013$. There was also a significant effect of essay version (for original, M = 0.0328; for revised, M = 0.0292), $F(1,335) = 8.84$, $p = 0.003$, $\eta^2_{partial} = 0.026$, and a marginal effect of feedback (for feedback, M = 0.0287; for no feedback, M = 0.0332), $F(1,335) = 3.686$, $p = 0.056$, $\eta^2_{partial} = 0.011$. Post hoc comparisons showed a significant difference between the original and revised essay for non-native speakers, $t(70) = 2.597$, $p = 0.011$, but no difference for native speakers, $t(267) = 1.883$, $p = 0.061$.

It should be noted that the effect sizes reported here are quite small and no doubt reflect a kind of 'floor effect' in these data. As discussed above, this rather proficient non-native group produced substantially fewer article errors than had been reported in studies of TOEFL examinees. In fact, their error rate was only slightly higher than the occasional, random errors of the native speakers. Despite this, statistically reliable effects were obtained.

We also split both the native and non-native groups by feedback condition (no feedback and feedback), and for each condition, we calculated how many students showed an improvement in their article error rate from the original first-draft to the revised, final version. Students whose error rate on the revised essay was within ± 0.005 of their original rate were classified as 'No Change'. Students with a lower error rate in their revised essay were classified as doing 'Better', while students who had a higher rate were classified as doing 'Worse'.

Table 4 shows the distribution of changes in article error rate across language groups and feedback conditions.

For native speakers, about one-third of the students showed improvement in both the no-feedback and feedback conditions; overall there was no significant difference in the distributions of 'Better', 'Worse', and 'No Change' between the conditions for the native speakers (χ^2 (2, N = 268) = 0.935, $p > 0.05$). By contrast, for the non-native speakers, the distribution in the feedback condition revealed a larger proportion of participants showing improvement as compared to the distribution for the no-feedback condition (χ^2 (2, N = 71) = 6.499, $p = 0.039$).

Table 4. Distribution of changes in article error rate by language group and feedback condition

Condition	Article error rate	Native speakers		Non-native speakers	
		Count	Proportion	Count	Proportion
No feedback	Better	29	**0.33**	8	**0.36**
	Worse	35	0.40	3	0.14
	No change	24	0.27	11	0.50
Feedback	Better	63	**0.35**	26	**0.53**
	Worse	61	0.34	13	0.27
	No change	56	0.31	10	0.20

The results of this study suggest that an article error detection module can reduce the article error rate in the writing of non-native speakers; however, it should be noted that this study is preliminary in the sense that it deals with only a single essay assignment over a period of just one week. In future work, we hope to track longitudinal changes in performance over many essays written by students over the course of many weeks or months.

Experiment 2 (MSR ESL Assistant): Do users distinguish between good and bad correction suggestions?

Background. The *ESL Assistant* is implemented as a web service in order to use large statistical resources, such as the Gigaword language model, which would not be available on a standard standalone desktop computer, and also to provide web-based search for example sentences. In addition to article and preposition error detection and correction, the *ESL Assistant* tackles a variety of other ESL errors (Gamon et al., 2009). In this section, however, we will focus entirely on the article and preposition modules. On a blind test corpus of web-scraped non-native writing from homepages and blogs, we achieve precision of 91% at 37% recall for articles. On the same data, preposition error detection is at 78% precision and 18% recall.

Procedure, data, and participants. A very simple user interface was implemented for the web service, with an additional option to download a Microsoft Office Outlook plugin that connects to the web service for the purpose of checking email text. Suggested corrections are indicated by a green squiggled line under a word in the input. If the user's mouse pointer hovers over the squiggled word, a list of suggestions is shown. Hovering over one of the suggestions will trigger a web search, as shown in Figure 2. Clicking on the suggestion will cause it to be substituted for the original input. The *ESL Assistant* was first deployed inside Microsoft for a trial period and was made available to the public in June 2008.

In both the internal and the external trials, user data were collected (with the user's permission) to assess site traffic and, more importantly, to serve as real user input so that system performance could be analyzed based on real data. Results reported in this section are from the public deployment, collected over a three-month period from 25 June 2008 through 24 September 2008.

The input was filtered to restrict the data to genuine user input. As with every publicly deployed web service, there are a number of users who just 'play' with the system, entering nonsense strings or copying text from the website itself and pasting it into the system. On the other hand, there are repeat users who use the system to check email text or other writing. For our evaluation, we analyzed only data from users who used *ESL Assistant* for at least four sessions.

There were 8714 sentences from 145 users and 863 user sessions that were retained for closer analysis. Manual analysis of the data revealed that the text could be grouped into six domains: email text (52%), general technical writing (26%), general non-technical writing (18%), essays (2%), printed media text (copied and pasted from published sources) (1%), and resumes (1%).

For the final analysis reported below, we used the email subset of the data since it most closely reflected a consistent use of the system on real data from a sizeable set of users. The email part consisted of 4550 sentences from 107 users and 702 sessions. While we do not log detailed information about users, we could get an approximate distribution among countries by looking at the user's machine locale. 46% of email users accessed the service from China, followed by 30% from the USA and 20% from Korea. Japan and Taiwan accounted for the remaining 4%.

The logged data also included all information about the suggestions made by the system, as well as the actions taken by the user. The user actions fell into four categories:

- No action: the user was presented with a suggestion in the form of a squiggled line under a word in the input, but took no action.
- Hover squiggle: the user hovered over the squiggle to have the suggestions displayed.
- Hover squiggle + hover suggestion: the user hovered over a squiggle, then hovered over a suggestion, triggering a web search.
- Hover squiggle + hover suggestion + accept suggestion: the user performed the whole series of actions from hovering over a squiggle to hovering over one or more suggestions to finally clicking on a suggested correction which led to a replacement of the original input with the suggested correction.

Results and discussion. The main question we would like to answer based on these data is whether the user's behavior resembles a blind acceptance of every suggested correction, or whether the users typically make careful choices in their acceptance of suggested corrections. The former would indicate that the usefulness of the service is very limited, since blind acceptance without reflection would have two undesirable consequences:

(a) It would introduce artificial errors into the user input (although system-centric evaluation numbers indicate that on average the number of errors in a text will decrease if all suggestions are accepted).
(b) It would not contribute to any learning on the part of the user.

As a first step towards answering this question, we examined the distribution of user actions and asked 'How often do users investigate suggestions and how often do they accept them?'

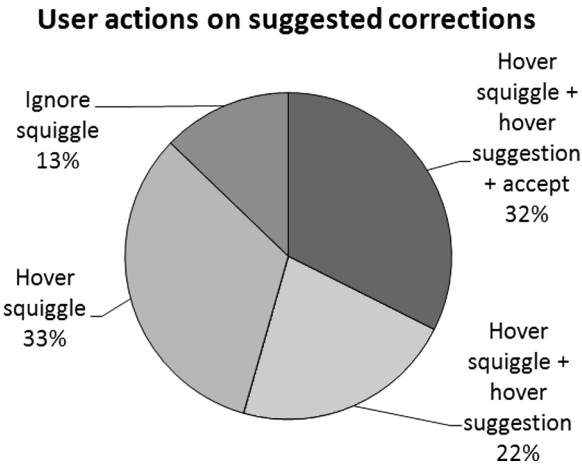

Figure 4. Actions taken by the user on suggested corrections

Of the 4550 email sentences, 12% triggered at least one article error and 8% triggered at least one preposition error. Figure 4 illustrates the distribution of user actions on the suggested corrections. In 13% of the cases, the suggested corrections were completely ignored by the user; in 33%, the user investigated the suggestions without further action; in 22%, the user hovered over one or more suggestions, triggering the web search component. Finally, in 32% of the cases, the user accepted the correction, causing it to be substituted for the originally entered string.

Based on these data, we believe it is safe to claim that the users indeed made selective decisions in their choice of action. The next question, however, is whether the user action led to an improvement in the entered text. In other words, did the users recognize valid suggestions and distinguish them from invalid ones? To answer this question, we performed an analysis of the accepted corrections where each accepted correction was categorized in one of the following ways:

- *Good*: the accepted correction fixed a problem in the user input.
- *Neutral*: the accepted suggestion was either a legitimate alternative of a well-formed original input, or the original input was ill-formed to the point where the suggested correction would neither improve nor further degrade the user input.
- *Bad*: the accepted suggestion resulted in an error or otherwise led to a degradation over the original user input.

Results for prepositions and articles are shown in Figure 5, and, for comparison, the baseline performance of the system on the same data is shown in Figure 6. The baseline performance indicates the distribution of all good/neutral/bad corrections that are provided by the system, as opposed to only those in the user-accepted corrections, as in Figure 5. The comparison illustrates that users favored valid suggestions over invalid ones. For articles, the overall rate of good suggestions was 62%, and increased to 70%

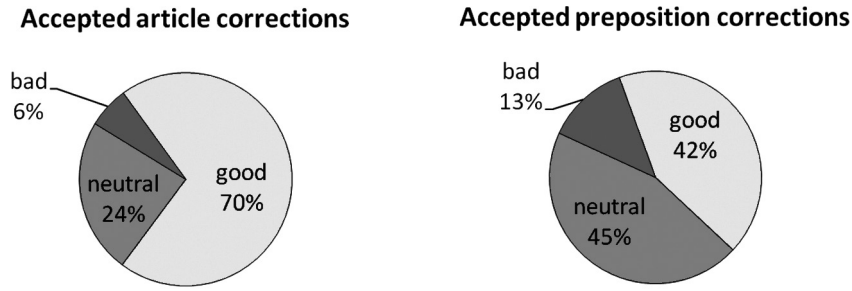

Figure 5. Evaluation of accepted article and preposition suggested corrections

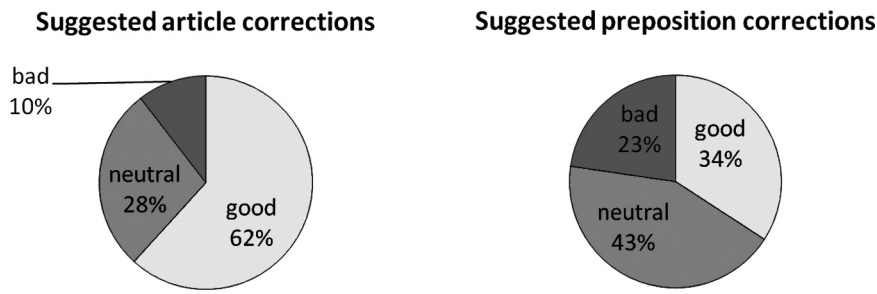

Figure 6. Evaluation of all article and preposition suggested corrections produced by the *ESL Assistant*

after the user made a decision. The system's false positive rate was 10%, which dropped to 6% after the user made a decision about accepting the suggestion. The difference in false positive rates was statistically significant: for the 34 users who accepted one or more suggested article corrections, Wilcoxon's signed ranks $T = 94$, $p < 0.001$. The situation was similar for preposition correction, where good corrections increased from 34% to 42% and false positives dropped from 23% to 13%. Here, too, the difference in false positive rates was statistically significant: for the 41 users who accepted one or more suggested corrections, Wilcoxon's signed ranks $T = 51$, $p < 0.001$.

The data collected during three months of our prototype trial deployment indicate that:

1. Users do not accept suggested corrections blindly; they are selective in their behavior.
2. Users make informed choices; they are able to distinguish correct suggestions from incorrect ones.

While the present study is only a first step in establishing the overall usefulness of an automated ESL proofing service, it does show that rather than making the decision for the user, which might lead to unreflective use of the service, the suggested corrections, with

alternatives, can aid the user in making an informed decision. One important question we would like to address in the future is the role that the web usage examples play in the user's behavior. In the present deployment of the web service, it is not possible to detect whether the users take the web search results into account in their decisions. Latencies between hovering over a suggestion and accepting a suggestion average 1.31 seconds, but the distribution of latencies is highly positively skewed, suggesting to us that at least some users have spent enough time that they may well have taken advantage of the web based examples. An answer to this question will likely require controlled user studies, possibly including automatic techniques, such as eye tracking, to closely examine the user's interaction with the system. Finally, the availability of real 'before-and-after' user data opens interesting avenues of future research. These data are hardly likely to be as reliable as manually annotated/corrected error corpora, but they are less expensive to collect and thus have the potential to surpass the manually annotated corpora in sheer volume and in total number of writers. As such, automatically collected user judgments can become a valuable resource for the tuning and training of error detection and correction systems.

Summary and Conclusion

In this paper, we have described some of the challenges that face language learners in mastering the usage of English articles and prepositions. We have also presented an overview of our computational systems for automatic error detection and correction. In the field of natural language processing, system-centric evaluation is the norm, but here we have focused instead on the user and have asked what effects feedback might have on the quality of the user's writing in a college essay and in email. There are, of course, some obvious shortcomings of the two experiments we report. Neither provides a longitudinal view of the effects that interacting with an error correction system may have had on the user. As a result, we are left to wonder if the benefits will extend beyond the current essay or the current email. Also lacking are more subjective data from the users to inform us of their attitudes and opinions about writing and the value of computer assistance, both before and after using the error correction system. As Heift and Schulze (2007) point out, attitudes are known to have an effect on the success of corrective feedback in the traditional language learning classroom and are likely to extend to the CALL environment as well. In future research, we hope to extend our work into both of these areas.

Even though they are limited in the ways described above, we believe that the experiments reported here have demonstrated the utility of article and preposition error correction systems. In a study of college writing using *Criterion*, feedback led to a significant increase in the proportion of non-native speakers who reduced their article error rates in their revised essays. In a study of email text, *ESL Assistant* provided suggested corrections for article and preposition errors plus actual example sentences from the web containing the suggested forms. The users' greater ability to recognize correct usage, even when they were unable to produce it, enabled them to select correct suggestions at a higher rate than the baseline performance rate of *ESL Assistant*.

Future developments in the computational modeling of language will almost certainly result in better systems for error detection and correction and in broader coverage of English syntax. We believe that, along with such improvements, it is important to continue

to focus on the value of the system's feedback to the user. By doing so, we can change the way system developers view the user's role from that of a passive recipient of information to that of an active participant in the processing of writing.

Acknowledgements

The authors wish to thank Claudia Leacock for her many substantive contributions to this work, Yoko Futagi for helpful comments on an earlier verison of this paper, and Ana Lipnevich for generously providing access to her data.

References

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.

Bitchener J, Young S, and Cameron D (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing, 14*(3), 191–205.

Burstein J, Chodorow M, and Leacock C (2004). Automated essay evaluation: The *Criterion* online writing service. *AI Magazine*, *25*(3), 27–36.

Chodorow M, Tetreault J, and Han N-R (2007). Detection of grammatical errors involving prepositions. *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*, 25–30.

Chapelle C (2001). *Computer applications in second language acquisition*. Cambridge: Cambridge University Press.

de Felice R, Pulman S (2007). Automatically acquiring models of preposition use. *Proceedings of the 4th Association for Computational Linguistics-Special Interest Group on Semantics (SIGSEM) Workshop on Prepositions*, 45–50.

Dalgish G (1985). Computer-assisted ESL research and courseware development. *Computers and Composition*, *2*(4), 45–62.

Diab N (1997). The transfer of Arabic in the English writings of Lebanese students. *The ESPecialist*, *18*(1), 71–83.

Eeg-Olofsson J, Knuttson O (2003). Automatic grammar checking for second language learners – the use of prepositions. *Proceedings of Nodalida*, Reykjavık, Iceland.

Gamon M, Gao J, Brockett C, Klementiev A, Dolan B, Belenko D, and Vanderwende L (2008). Using contextual speller techniques and language modeling for ESL error correction. *Proceedings of IJCNLP*, Hyderabad, India.

Gamon M, Leacock C, Brockett C, Dolan B, Gao J, Belenko D, and Klementiev A (2009): Using statistical techniques and web search to correct ESL errors. *CALICO Journal,* special edition on 'Automatic Analysis of Learner's Language'.

Gao J, Goodman J, and Miao J (2001). The use of clustering techniques for language modeling – application to Asian languages. *Computational Linguistics and Chinese Language Processing*, *6*(1), 27–60.

Golding A, Roth D, Mooney J, and Cardie C (1999). A winnow based approach to context-sensitive spelling correction. *Machine Learning*, 107–130.

Han N-R, Chodorow M, and Leacock C (2006). Detecting errors in English article usage by non-native speakers. *Natural Language Engineering, 12*(2), 115–129.

Heift T (2004). Corrective feedback and learner uptake in CALL. *ReCALL*, *16*(2), 416–431.

Heift T, Schulze M (2007). *Errors and intelligence in computer-assisted language learning: Parsers and pedagogues.* New York: Routledge.

Izumi E, Uchimoto K, Saiga T, Supnithi T, and Ishara H (2003). Automatic error detection in the Japanese learner's English spoken data. *Companion Volume to the Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Linguistic Data Consortium (LDC). (2003). English Gigaword corpus.

Lipnevich A, Smith J (2009) Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied*. 15(4): 319–333.

Nagata R, Kawai A, Morihiro K, and Isu N (2006). A feedback-augmented method for detecting errors in the writing of learners of English. *Proceedings of the ACL/COLING.*

Nguyen P, Gao J, and Mahajan M (2007). MSRLM: A scalable language modeling toolkit. *Microsoft Research Technical Report, MSR-TR-2007-144*.

Ratnaparkhi A (1997). A simple introduction to maximum entropy models for natural language processing. *Technical Report IRCS Report 97-08, Institute for Research in Cognitive Science*, Philadelphia, PA.

Tetreault J, Chodorow M (2008a). The ups and downs of preposition error detection. *COLING*, Manchester, UK.

Tetreault J, Chodorow M (2008b). Native judgments of non-native usage: Experiments in preposition error detection. *COLING Workshop on Human Judgments in Computational Linguistics*, Manchester, UK.

Turner J, Charniak E (2007). Language modeling for determiner selection. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, 177–180.