# User Input and Interactions on *Microsoft Research ESL Assistant*

**Claudia Leacock**
Butler Hill Group
P.O. Box 935
Ridgefield, CT, 06877, USA
claudia.leacock@gmail.com

**Michael Gamon**
Microsoft Research
One Microsoft Way
Redmond, WA, 98052, USA
mgamon@microsoft.com

**Chris Brockett**
Microsoft Research
One Microsoft Way
Redmond, WA, 98052, USA
chrisbkt@microsoft.com

## Abstract

*ESL Assistant* is a prototype web-based writing-assistance tool that is being developed for English Language Learners. The system focuses on types of errors that are typically made by non-native writers of American English. A freely-available prototype was deployed in June 2008. User data from this system are manually evaluated to identify writing domain and measure system accuracy. Combining the user log data with the evaluated rewrite suggestions enables us to determine how effectively English language learners are using the system, across rule types and across writing domains. We find that repeat users typically make informed choices and can distinguish correct suggestions from incorrect.

## 1 Introduction

Much current research in grammatical error detection and correction is focused on writing by English Language Learners (ELL). The *Microsoft Research ESL Assistant* is a web-based proofreading tool designed primarily for ELLs who are native speakers of East-Asian languages. Initial system development was informed by pre-existing ELL error corpora, which were used both to identify common ELL mistakes and to evaluate system performance. These corpora, however, were created from data collected under arguably artificial classroom or examination conditions, leaving unresolved the more practical question as to whether the *ESL Assistant* can actually help a person who produced the text to improve their English language writing skills in course of more realistic everyday writing tasks.

In June of 2008, a prototype version of this system was made freely available as a web service[1]. Both the writing suggestions that visitors see and the actions that they then take are recorded. As these more realistic data begin to accumulate, we can now begin to answer the above question.

## 2 Related Work

Language learner error correction techniques typically fall into either of two categories: rule-based or data-driven. Eeg-Olofsson and Knutsson (2003) report on a rule-based system that detects and corrects preposition errors in non-native Swedish text. Rule-based approaches have also been used to predict definiteness and indefiniteness of Japanese noun phrases as a preprocessing step for Japanese to English machine translation (Murata and Nagao 1993; Bond et al, 1994; Heine, 1998), a task that is similar to the prediction of English articles. More recently, data-driven approaches have gained popularity and been applied to article prediction in English (Knight and Chander 1994; Minnen et al, 2000; Turner and Charniak 2007), to an array of Japanese learners' errors in English (Izumi et al, 2003), to verb errors (Lee and Seneff, 2008), and to article and preposition correction in texts written by non-native ELLs (Han et al, 2004, 2006; Nagata et al, 2005; Nagata et al, 2006; De Felice and Pulman, 2007; Chodorow et al, 2007; Gamon et al, 2008, 2009; Tetreault and Chodorow, 2008a).

[1] http://www.eslassistant.com

| | | |
|---|---|---|
| **Noun Related (61%)** | Articles *(ML)* | We have just checked **the* our stock.<br>life is **journey/a journey*, travel it well!<br>I think it 's **a/the* best way to resolve issues like this. |
| | Noun Number | London is one of the most attractive **city/cities* in the world.<br>You have to write down all the details of each **things/thing* to do.<br>Conversion always takes a lot of **efforts/effort*. |
| | Noun Of Noun | Please send the **feedback of customer/customer feedback* to me by mail. |
| **Preposition Related (27%)** | Preposition *(ML)* | I'm **on* home today, call me if you have a problem.<br>It seems ok and I did not pay much attention **on/to* it.<br>Below is my contact, looking forward **your/to your* response, thanks! |
| | Verb and Preposition | Ben is involved **this/in* this transaction.<br>I should **to ask/ask* a rhetorical question …<br>But I'll think **it/about it* a second time. |
| **Verb Related (10%)** | Gerund / Infinitive *(ML)* | He got me **roll/to roll* up my sleeve and make a fist.<br>On Saturday, I with my classmate went **eating/to eat*.<br>After **get/getting* a visa, I want to study in New York. |
| | Auxiliary Verb *(ML)* | To learn English we should **be speak/speak* it as much as possible .<br>Hope you will **happy/be happy* in Taiwan .<br>what **is/do* you want to say? |
| | Verb formation | If yes, I will **attached/attach* and resend to Geoff .<br>The time and setting are **display/displayed* at the same time.<br>You had **order/ordered* 3 items … this time.<br>I am really **hope/hoping* to visit UCLA. |
| | Cognate/Verb Confusion | We cannot **image/imagine* what the environment really is at the site of end user . |
| | Irregular Verbs | I **teached/taught* him all the things that I know … |
| **Adj Related (2%)** | Adjective Confusions | She is very **interesting/interested* in the problem.<br>So **Korea/Korean* Government is intensively fostering trade .<br>… and it is **much/much more* reliable than your Courier Service. |
| | Adjective order | Employing the **Chinese ancient/ancient Chinese* proverb, that is … |

Table 1: ESL Assistant grammatical error modules. *ML* modules are machine learned.

## 3 ESL Assistant

*ESL Assistant* takes a hybrid approach that combines statistical and rule-based techniques. Machine learning is used for those error types that are difficult to identify and resolve without taking into account complex contextual interactions, like article and preposition errors. Rule-based approaches handle those error types that are amenable to simpler solutions. For example, a regular expression is sufficient for identifying when a modal is (incorrectly) followed by a tensed verb.

The output of all modules, both machine-learned and rule-based, is filtered through a very large language model. Only when the language model finds that the likelihood of the suggested rewrite is sufficiently larger than the original text is a suggestion shown to the user. For a detailed description of *ESL Assistant's* architecture, see Gamon et al (2008, 2009).

Although this and the systems cited in section 2 are designed to be used by non-native writers, system performance is typically reported in relation to native text – the prediction of a preposition, for example, will ideally be consistent with usage in native, edited text. An error is counted each time the system predicts a token that differs from the observed usage and a correct prediction is counted each time the system predicts the usage that occurs in the text. Although somewhat artificial, this approach to evaluation offers the advantages of being fully automatable and having abundant quantities

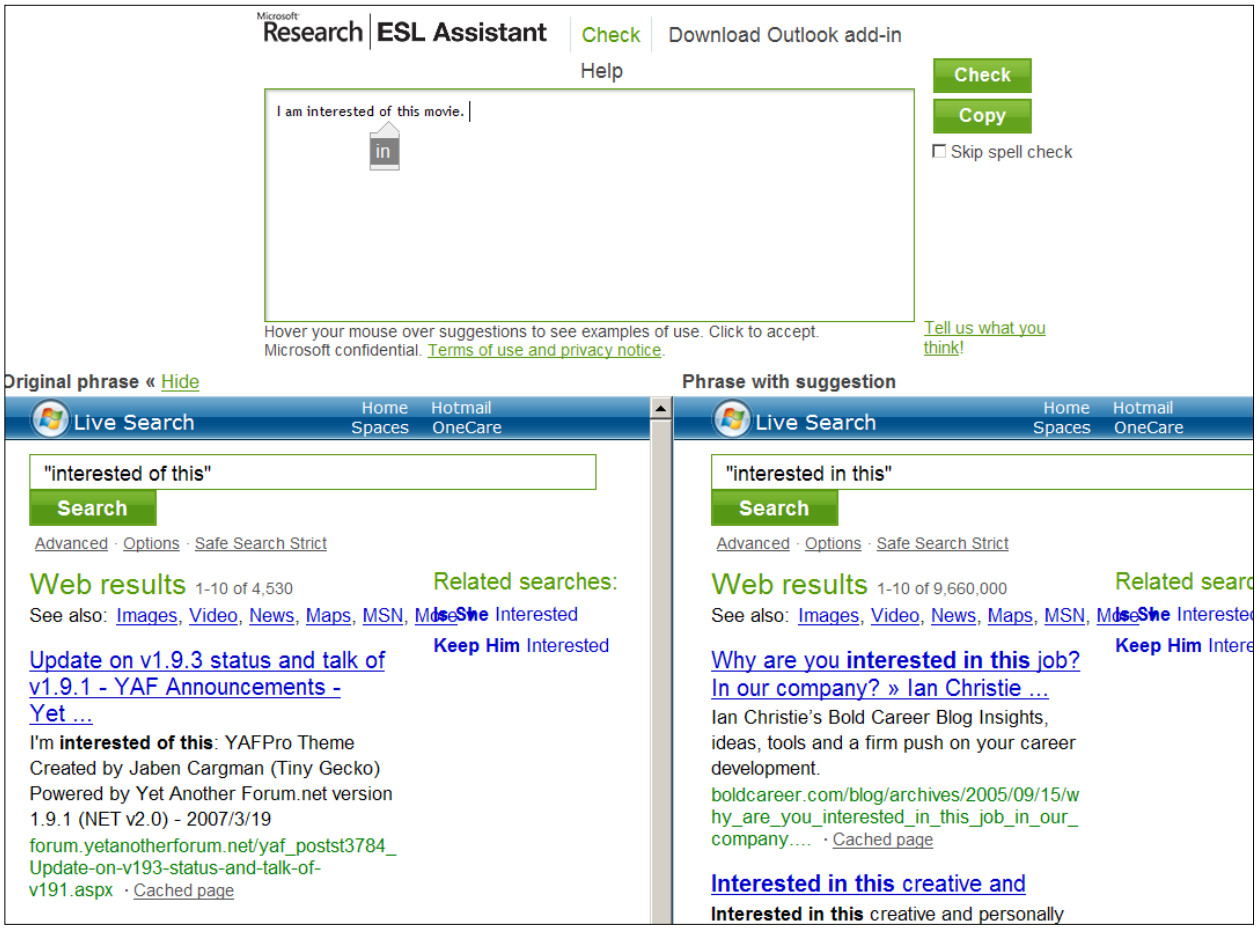


Figure 1: Screen shot of *ESL Assistant*

of edited data readily available. With respect to prepositions and articles, the *ESL Assistant's* classifiers achieve state-of-the-art performance when compared to results reported in the literature (Gamon et al, 2008), inasmuch as comparison is possible when the systems are evaluated on different samples of native text. For articles, the system had 86.76% accuracy as compared to 86.74% reported by Turner and Charniak (2007), who have the most recently reported results. For the harder problem of prepositions, *ESL Assistant's* accuracy is comparable to those reported by Tetreault and Chodorow (2008a) and De Felice and Pulman (2007).

### 3.1 Error Types

The ELL grammatical errors that *ESL Assistant* tries to correct were distilled from analysis of the most frequent errors made in Chinese and Japanese English language learner corpora (Gui and Yang, 2001; Izumi et al. 2004). The error types are shown in Table 1: modules identified with *ML* are machine-learned, while the remaining modules are rule-based. *ESL Assistant* does not attempt to identify those errors currently found by Microsoft Word™, such as subject/verb agreement.

*ESL Assistant* further contains a component to help address lexical selection issues. Since this module is currently undergoing major revision, we will not report on the results here.

### 3.2 System Development

Whereas evaluation on native writing is essential for system development and enables us to compare *ESL Assistant* performance with that of other reported results, it tells us little about how the system would perform when being used by its true target audience – non-native speakers of English engaged in real-life writing tasks. In this context, performance measurement inevitably entails manual evaluation, a process that is notoriously time consuming, costly and potentially error-prone. Human inter-rater agreement is known to be problematic

on this task: it is likely to be high in the case of certain user error types, such as over-regularized verb inflection (where the system suggests replacing "writed" with "wrote"), but other error types are difficult to evaluate, and much may hinge upon who is performing the evaluation: Tetreault and Chodorow (2008b) report that for the annotation of preposition errors "using a single rater as a gold standard, there is the potential to over- or underestimate precision by as much as 10%."

With these caveats in mind, we employed a single annotator to evaluate system performance on native data from the 1-million-word Chinese Learner's of English corpus (Gui and Yang, 2001; 2003). Half of the corpus was utilized to inform system development, while the remaining half was held back for "unseen" evaluation. While the absolute numbers for some modules are more reliable than for others, the relative change in numbers across evaluations has proven a beneficial yardstick of improved or degraded performance in the course of development.

### 3.3 The User Interface and Data Collection

Figure 1 shows the *ESL Assistant* user interface. When a visitor to the site types or pastes text into the box provided and clicks the "Check" button, the text is sent to a server for analysis. Any locations in the text that trigger an error flag are then displayed as underscored with a wavy line (known as a "squiggle"). If the user hovers the mouse over a squiggle, one or more suggested rewrites are displayed in a dropdown list. Then, if the user hovers over one of these suggestions, the system launches parallel web searches for both original and rewrite phrases in order to allow the user to compare real-word examples found on the World Wide Web. To accept a suggestion, the user clicks on the suggested rewrite, and the text is emended. Each of these actions, by both system and user, are logged on the server.

Since being launched in June, 2008, *ESL Assistant* has been visited over 100,000 times. Currently, the web page is being viewed between one to two thousand times every day. From these numbers alone it seems safe to conclude that there is much public interest in an ESL proofreading tool.

Fifty-three percent of visitors to the *ESL Assistant* web site are from countries in East Asia – its primary target audience – and an additional 15% are from the United States. Brazil, Canada, Germany, and the United Kingdom each account for about 2% of the site's visitors. Other countries represented in the database each account for 1% or less of all those who visit the site.

### 3.4 Database of User Input

User data are collected so that system performance can be evaluated on actual user input – as opposed to running pre-existing learner corpora through the system. User data provide invaluable insight into which rewrite suggestions users spend time viewing, and what action they subsequently take on the basis of those suggestions.

These data must be screened, since not all of the textual material entered by users in the web site is valid learner English language data. As with any publicly deployed web service, we find that numerous users will play with the system, entering nonsense strings or copying text from elsewhere on the website and pasting it into the text box.

To filter out the more obvious non-English data, we eliminate input that contains, for example, no alphabetic characters, no vowels/consonants in a sentence, or no white space. "Sentences" consisting of email subject lines are also removed, as are all the data entered by the *ESL Assistant* developers themselves. Since people often enter the same sentence many times within a session, we also remove repetitions of identical sentences within a single session.

Approximately 90% of the people who have visited the web site visit it once and never return. This behavior is far from unusual on the web, where site visits may have no clear purpose beyond idle curiosity. In addition, some proportion of visitors may in reality be automated "bots" that can be nearly indistinguishable from human visitors.

Nevertheless, we observe a significant number of repeat visitors who return several times to use the system to proofread email or other writing, and these are the users that we are intrinsically interested in. To measure performance, we therefore decided to evaluate on data collected from users who logged on and entered plausibly English-like text on at least four occasions. As of 2/10/2009, the frequent user database contained 39,944 session-unique sentences from 578 frequent users in 5,305 sessions.

Data from these users were manually annotated to identify writing domains as shown in Table 2. Fifty-three percent of the data consists of people proofreading email.[2] The dominance of email data is presumably due to an Outlook plug-in that is available on the web site, and automates copying email content into the tool. The non-technical domain consists of student essays, material posted on a personal web site, or employees writing about their company – for example, its history or processes. The technical writing is largely conference papers or dissertations in the fields of, for example, medicine and computer science. The "other" category includes lists and resumes (a writing style that deliberately omits articles and grammatical subjects), as well as text copied from online newspapers or other media and pasted in.

| Writing Domain | Percent |
|---|---|
| Email | 53% |
| Non-technical / essays | 24% |
| Technical / scientific | 14% |
| Other (lists, resumes, etc) | 4% |
| Unrelated sentences | 5% |

Table 2: Writing domains of frequent users

Sessions categorized as "unrelated sentences" typically consist of a series of short, unrelated sentences that each contain one or more errors. These users are testing the system to see what it does. While this is a legitimate use of any grammar checker, the user is unlikely to be proofreading his or her writing, so these data are excluded from evaluation.

## 4 System Evaluation & User Interactions

We are manually evaluating the rewrite suggestions that *ESL Assistant* generated in order to determine both system accuracy and whether user acceptances led to an improvement in their writing. These categories are shown in Table 3. Note that results reported for non-native text look very different from those reported for native text (discussed in Section 3) because of the *neutral* categories which do not appear in the evaluation of native text. Systems reporting 87% accuracy on native text cannot achieve anything near that on non-native ELL text because almost one third of the flags fall into a neutral category.

| Evaluation | Subcategory: Description |
|---|---|
| **Good** | **Correct flag**: The correction fixes a problem in the user input. |
| **Neutral** | **Both Good**: The suggestion is a legitimate alternative to well-formed original input: *I like working/to work.* |
| | **Misdiagnosis**: the original input contained an error but the suggested rewrite neither improves nor further degrades the input: *If you ~~have~~ fail machine on hand.* |
| | **Both Wrong**: An error type is correctly diagnosed but the suggested rewrite does not correct the problem: *can you give me suggestion.* (suggests *the* instead of *a*) |
| | **Non-ascii:** A non-ascii or text markup character is in the immediate context. |
| **Bad** | **False Flag**: The suggestion resulted in an error or would otherwise lead to a degradation over the original user input. |

Table 3: Evaluation categories

In 51% of the 39,944 frequent user sentences, the system generated at least one grammatical error flag, for a total of 17,832 flags. Thirteen percent of the time, the user ignored the flags. The remaining 87% of the flags were inspected by the user, and of those, the user looked at the suggested rewrites without taking further action 31% of the time. For 28% of the flags, the user hovered over a suggestion to trigger a parallel web search but did not accept the proposed rewrite. Nevertheless, 41% of inspected rewrites were accepted, causing the original string in the text to be revised. Overall, the users inspected about 15.5K suggested rewrites to accept about 6.4K. A significant number of users appear to be inspecting the suggested revisions and making deliberate choices to accept or not accept.

The next question is: Are users making the right choices? To help answer this question, 34% of the user sessions have been manually evaluated for system accuracy – a total of approximately 5.1K grammatical error flags. For each error category and for the three major writing domains, we:

[2] These are anonymized to protect user privacy.

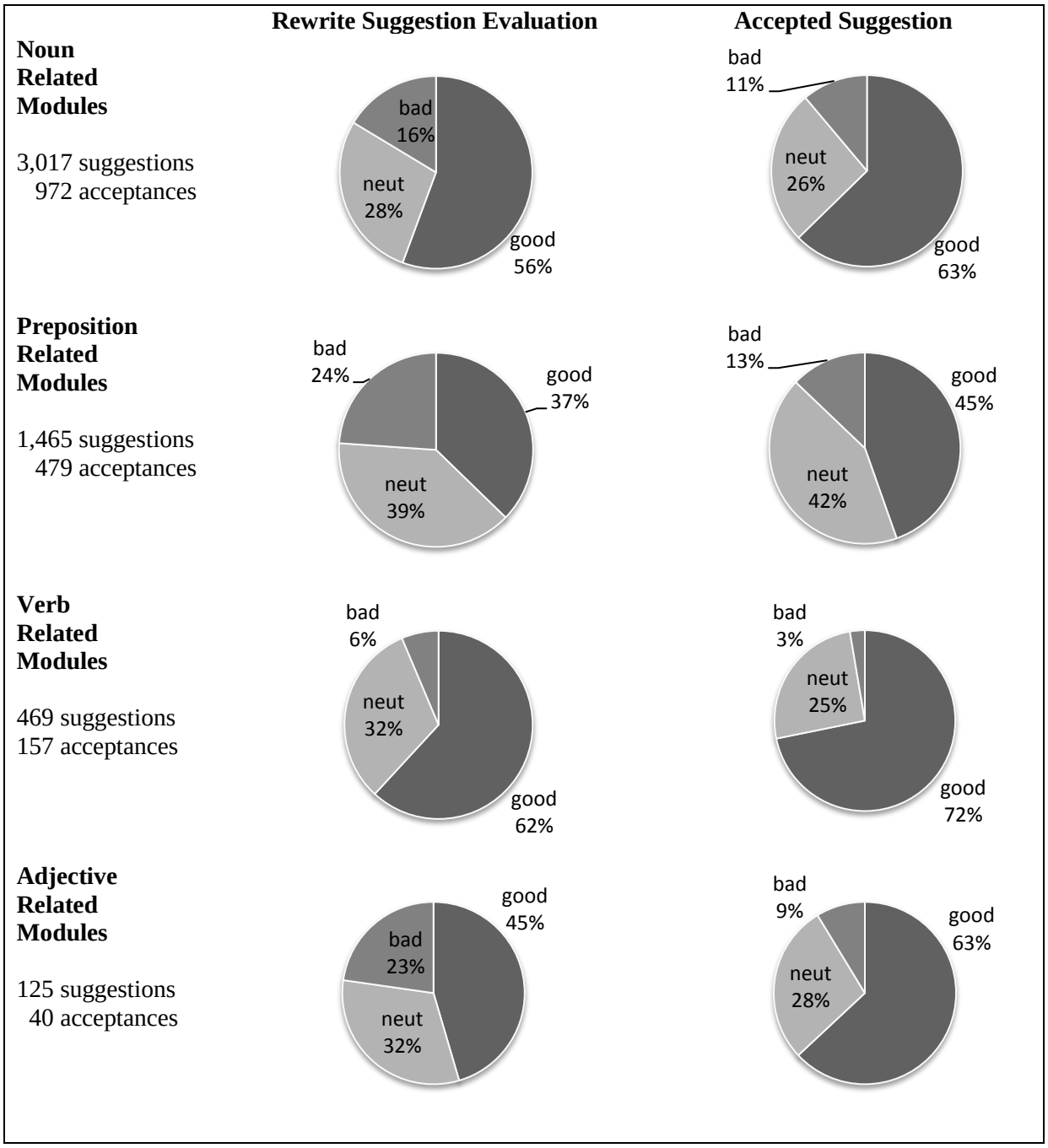


Figure 2: User interactions by module category

1. Calculated system accuracy for all flags, regardless of user actions.
2. Calculated system accuracy for only those rewrites that the user accepted
3. Compared the ratio of good to bad flags.

Results for the individual error categories are shown in Figure 2. Users consistently accept a greater proportion of good suggestions than they do bad ones across all error categories. This is most pronounced for the adjective-related modules, where the overall rate of good suggestions improved 17.6% after the user made the decision to accept a suggestion, while the system's false positive rate dropped 14.1% after the decision. For the noun-related modules, the system's most produc-

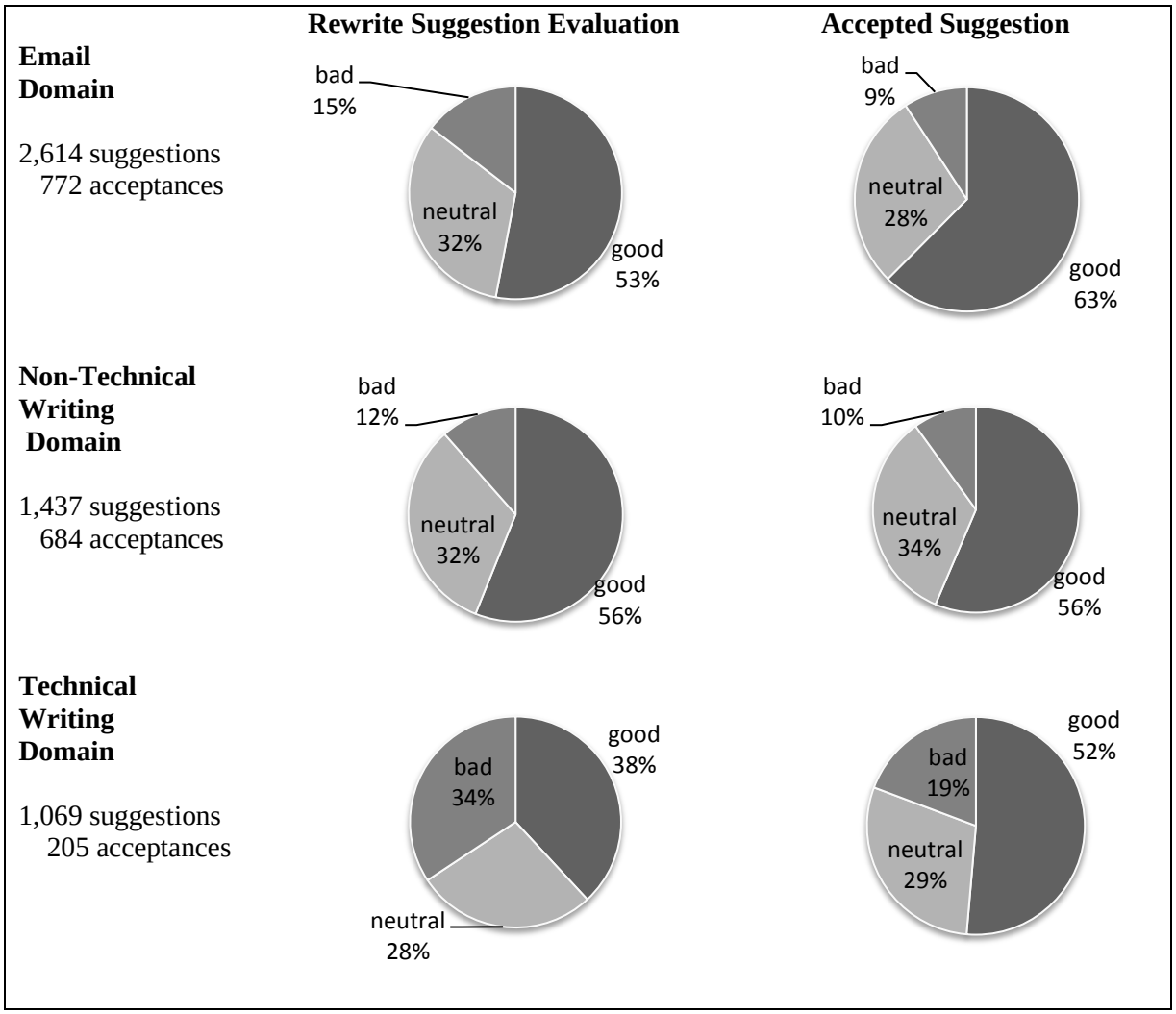


Figure 3: User interactions by writing domain

tive modules, the overall good flag rate increased by 7% while the false positive rate dropped 5%. All differences in false positive rates are statistically significant in Wilcoxon's signed-ranks test.

When all of the modules are evaluated across the three major writing domains, shown in figure 3, the same pattern of user discrimination between good and bad flags holds. This is most evident in the technical writing domain, where the overall rate of good suggestions improved 13.2% after accepting the suggestion and the false positive rate dropped 15.1% after the decision. It is least marked for the essay/nontechnical writing domain. Here the overall good flag rate increased by only .3% while the false positive rate dropped 1.6%. Again, all of the differences in false positive rates are statistically significant in Wilcoxon's signed-ranks test. These findings are consistent with those for the machine learned articles and prepositions modules in the email domain (Chodorow et al, *under review*).

A probable explanation for the differences seen across the domains is that those users who are proofreading non-technical writing are, as a whole, less proficient in English than the users who are writing in the other domains. Users who are proofreading technical writing are typically writing a dissertation or paper in English and therefore tend to relatively fluent in the language. The email domain comprises people who are confident enough in their English language skills to communicate with colleagues and friends by email in English. With the essay/non-technical writers, it often is not clear who the intended audience is. If there *is* any indication of audience, it is often an instructor. Users in this domain appear to be the least English-

language proficient of the *ESL Assistant* users, so it is unsurprising that they are less effective in discriminating between good and bad flags than their more proficient counterparts. Thus it appears that those users who are most in need of the system are being helped by it least – an important direction for future work.

Finally, we look at whether the neutral flags, which account for 29% of the total flags, have any effect. The two neutral categories highlighted in Table 3, flags that either misdiagnose the error or that diagnose it but do not correct it, account for 74% of *ESL Assistant's* neutral flags. Although these suggested rewrites do not improve the sentence, they do highlight an environment that contains an error. The question is: What is the effect of identifying an error when the rewrite doesn't improve the sentence?

To estimate this, we searched for cases where *ESL Assistant* produced a neutral flag and, though the user did not accept the suggestion, a revised sentence that generated no flag was subsequently submitted for analysis. For example, one user entered: "This early morning i got a from head office …". *ESL Assistant* suggested deleting *from*, which does not improve the sentence. Subsequently, in the same session, the user submitted, "This early morning I heard from the head office …" In this instance, the system correctly identified the location of an error. Moreover, even though the suggested rewrite was not a good solution, the information was sufficient to enable the user to fix the error on his or her own.

Out of 1,349 sentences with neutral suggestions that were not accepted, we identified (using a fuzzy match) 215 cases where the user voluntarily modified the sentence so that it contained no flag, without accepting the suggestion. In 44% of these cases, the user had simply typed in the suggested correction instead of accepting it – indicating that true acceptance rates might be higher than we originally estimated. Sixteen percent of the time, the sentence was revised but there remained an error that the system failed to detect. In the other 40% of cases, the voluntary revision improved the sentence. It appears that merely pointing out the possible location of an error to the user is often sufficient to be helpful.

## 5 Conclusion

In conclusion, judging from the number of people who have visited the *ESL Assistant* web site, there is considerable interest in ESL proofreading tools and services.

When using the tool to proofread text, users do not accept the proposed corrections blindly – they are selective in their behavior. More importantly, they are making informed choices – they can distinguish correct suggestions from incorrect ones. Sometimes identifying the location of an error, even when the solution offered is wrong, itself appears sufficient to cause the user to repair a problem on his or her own. Finally, the user interactions that we have recorded indicate that current state-of-the-art grammatical error correction technology has reached a point where it can be helpful to English language learners in real-world contexts.

## Acknowledgments

We thank Bill Dolan, Lucy Vanderwende, Jianfeng Gao, Alexandre Klementiev and Dmitriy Belenko for their contributions to the *ESL Assistant* system. We are also grateful to the two reviewers of this paper who provided valuable feedback.

## References

Francis Bond, Kentaro Ogura, and Satoru Ikehara. 1994. Countability and number in Japanese to English machine translation. In *Proceedings of the 15th Conference on Computational Linguistics* (pp. 32-38). Kyoto, Japan.

Martin Chodorow, Michael Gamon, and Joel Tetreault. Under review. The utility of grammatical error detection systems for English language learners: Feedback and Assessment.

Martin Chodorow, Joel Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions* (pp. 25-30). Prague, Czech Republic.

Rachele De Felice and Stephen G. Pulman. 2007. Automatically acquiring models of preposition use. In *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions* (pp. 45-50). Prague, Czech Republic.

Jens Eeg-Olofsson and Ola Knutsson. 2003. Automatic grammar checking for second language learners – the use of prepositions. *Proceedings of NoDaLiDa* 2003. Reykjavik, Iceland.

Michael Gamon, Jianfeng Gao, Chris Brockett, Alexandre Klementiev, William B. Dolan, Dmitriy Belenko, and Lucy Vanderwende. 2008. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of the Third International Joint Conference on Natural Language Processing* (pp. 449-455). Hyderabad, India.

Michael Gamon, Claudia Leacock, Chris Brockett, William B. Dolan, Jianfeng Gao, Dmitriy Belenko, and Alexandre Klementiev. 2009. Using statistical techniques and web search to correct ESL errors. To appear in *CALICO Journal, Special Issue on Automatic Analysis of Learner Language*.

Shicun Gui and Huizhong Yan. 2001. Computer analysis of Chinese learner English. Presentation at Hong Kong University of Science and Technology.http://lc.ust.hk/~centre/conf2001/keynote/subsect4/yang.pdf.

Shicun Gui and Huizhong Yang. (Eds.). 2003. *Zhongguo Xuexizhe Yingyu Yuliaohu. (Chinese Learner English Corpus)*. Shanghai Waiyu Jiaoyu Chubanshe. (In Chinese).

Na-Rae Han, Martin Chodorow, and Claudia Leacock 2004). Detecting errors in English article usage with a maximum entropy classifier trained on a large, diverse corpus. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal.

Na-Rae Han, Martin Chodorow, and Claudia Leacock 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering,* 12**(**2), 115-129.

Julia E. Heine. 1998. Definiteness predictions for Japanese noun phrases. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (pp. 519-525). Montreal, Canada.

Emi Izumi, Kiyotaka Uchimoto, Toyomi Saiga, Thepchai Supnithi, and Hitoshi Isahara. 2003. Automatic error detection in the Japanese learners' English spoken data. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 145-148). Sapporo, Japan.

Kevin Knight and Ishwar Chander,. 1994. Automatic postediting of documents. In *Proceedings of the 12th National Conference on Artificial Intelligence* (pp. 779-784). Seattle: WA.

John Lee. 2004. Automatic article restoration. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 31-36). Boston, MA.

John Lee and Stephanie Seneff. 2008. Correcting misuse of verb forms. In *Proceedings of ACl-08/HLT* (pp. 174-182). Columbus, OH.

Guido Minnen, Francis Bond, and Anne Copestake. 2000. Memory-based learning for article generation. In *Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop* (pp. 43-48). Lisbon, Portugal.

Masaki Murata and Makoto Nagao. 1993. Determination of referential property and number of nouns in Japanese sentences for machine translation into English. In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation* (pp. 218-225). Kyoto, Japan.

Ryo Nagata, Takahiro Wakana, Fumito Masui, Atsui Kawai, and Naoki Isu. 2005. Detecting article errors based on the mass count distinction. In R. Dale, W. Kam-Fie, J. Su and O.Y. Kwong (Eds.) *Natural Language Processing - IJCNLP 2005, Second International Joint Conference Proceedings* (pp. 815-826). New York: Springer.

Ryo Nagata, Atsuo Kawai, Koichiro Morihiro, and Naoki Isu. 2006. A feedback-augmented method for detecting errors in the writing of learners of English. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (pp. 241-248). Sydney, Australia.

Joel Tetreault and Martin Chodorow. 2008a. The ups and downs of preposition error detection in ESL. COLING. Manchester, UK.

Joel Tetreault and Martin Chodorow. 2008b. Native judgments of non-native usage: Experiments in preposition error detection. In *Proceedings of the Workshop on Human Judgments in Computational Linguistics, 22nd International Conference on Computational Linguistics* (pp 43-48). Manchester, UK.

Jenine Turner and Eugene Charniak. 2007. Language modeling for determiner selection. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*; *Companion Volume, Short Papers* (pp. 177-180). Rochester, NY.