

Does Automated Feedback in a Proofreading Tool Help an English Language Learner?

Claudia Leacock, Butler Hill Group

Michael Gamon, Microsoft Research

Chris Bocket, Microsoft Research

... and

William B. Dolan, Microsoft Research

Jianfeng Gao, Microsoft Research

Dmitriy Belenko, Microsoft Research

Lucy Vanderwende, Microsoft Research

Alexandre Klementiev, University of Illinois
at Urbana Champaign

ESL Assistant

- March 2008: CALICO Workshop: Gamon et al.
 - System Description & Evaluation. No user action.
 - System performance is state-of-the-art
- June 24, 2008: ESL Assistant goes live!
- 2009 CALICO Workshop Presentation
 - System Usage
 - Evaluation
 - User Interactions: What they saw. What they did.

Most frequent errors made by East Asian non-native speakers

Noun Related: Articles (inclusion & choice), Noun Number, Noun of Noun

- I think it's ***a/the** best way to resolve issues like this.
- Conversion always takes a lot of ***efforts/effort**.
- Please send the ***feedback of customer/customer feedback** to me by mail.

Preposition Related: inclusion & choice

- It seems ok and I did not pay much attention ***on/to** it.
- I should ***to ask/ask** a rhetorical question.

Verb Related: Gerund/Infinitive Confusion, Auxiliary Verb Error, Verb Formation Errors (6), Cognate/ Verb confusion, Irregular Verbs

- On Saturday, I with my classmate went ***eating/to eat**.
- Hope you will ***happy/be happy** in Taiwan.
- I ***tached/taught** him all the things I know.

Adjective Related: Adjective Confusion (4), Adjective Order

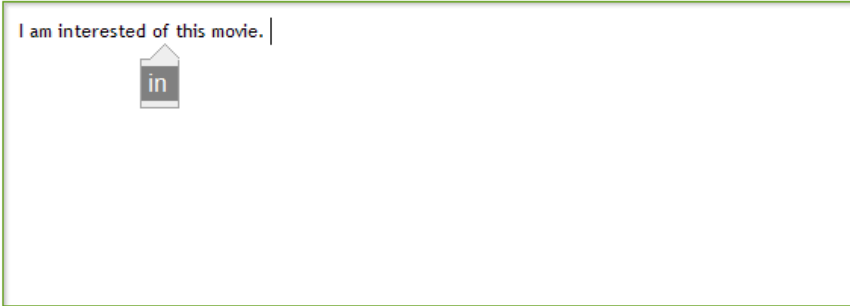
- She is very ***interesting/interested** in the problem.
- So ***Korea/Korean Government** is intensely fostering trade.

Users and Data Collection

ESL Assistant User Interface

Microsoft Research | **ESL Assistant** | [Check](#) | [Download Outlook add-in](#)
[Help](#)

I am interested of this movie. |



Skip spell check

[Check](#)
[Copy](#)

[Tell us what you think!](#)

Hover your mouse over suggestions to see examples of use. Click to accept.
Microsoft confidential. [Terms of use and privacy notice.](#)

Original phrase « [Hide](#)

Live Search Home Spaces Hotmail OneCare

"interested of this"

[Search](#)

[Advanced](#) · [Options](#) · [Safe Search Strict](#)

Web results 1-10 of 4,530

See also: [Images](#), [Video](#), [News](#), [Maps](#), [MSN](#), [More](#)

[Is She Interested](#)
[Keep Him Interested](#)

[Update on v1.9.3 status and talk of v1.9.1 - YAF Announcements - Yet ...](#)

I'm **interested of this**: YAFPro Theme
Created by Jaben Cargman (Tiny Gecko)
Powered by Yet Another Forum.net version 1.9.1 (NET v2.0) - 2007/3/19
[forum.yetanotherforum.net/yaf_postst3784_Update-on-v193-status-and-talk-of-v191.aspx](#) · [Cached page](#)

Phrase with suggestion

Live Search Home Spaces Hotmail OneCare

"interested in this"

[Search](#)

[Advanced](#) · [Options](#) · [Safe Search Strict](#)

Web results 1-10 of 9,660,000

See also: [Images](#), [Video](#), [News](#), [Maps](#), [MSN](#), [More](#)

[Is She Interester](#)
[Keep Him Intere](#)

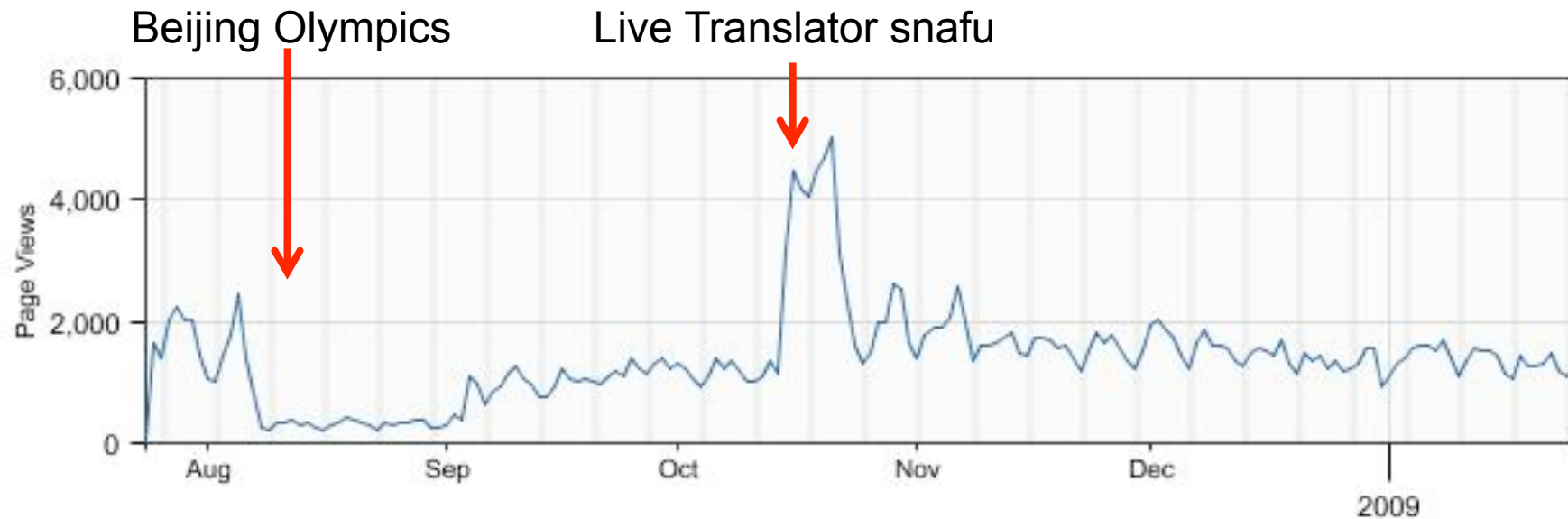
[Why are you interested in this job? In our company? » Ian Christie ...](#)
Ian Christie's Bold Career Blog Insights, ideas, tools and a firm push on your career development.
[boldcareer.com/blog/archives/2005/09/15/why_are_you_interested_in_this_job_in_our_company....](#) · [Cached page](#)

[Interested in this creative and](#)
Interested in this creative and personally

Page Views per Day

Traffic via
website links:

Windows Live Translator	35%
Chinese MSN	13%
Taiwan MSN	11%
Korean MSN	7%



Page Views

Thu. 24 Jul. 2008 - Sun. 25 Jan. 2009

■ Selected Period

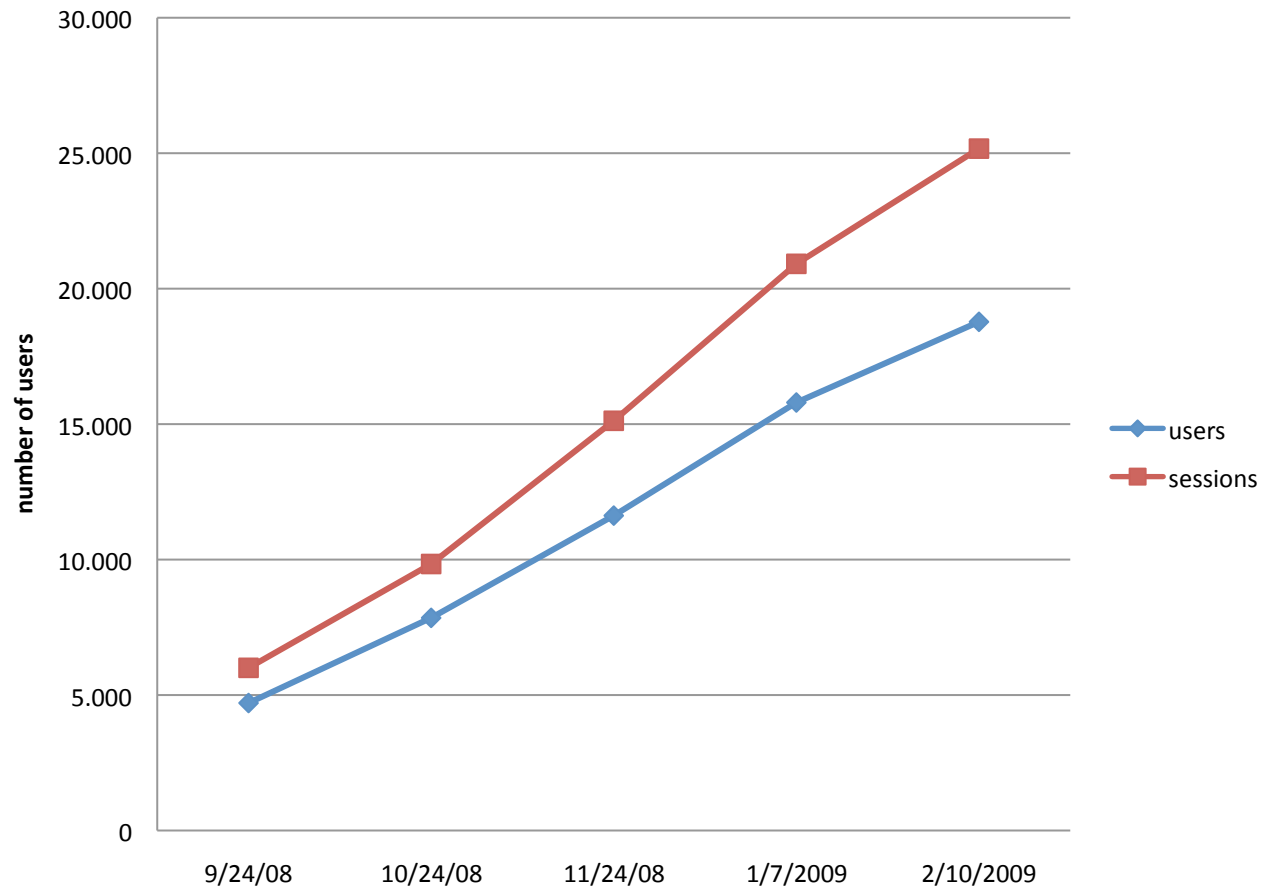
User Location



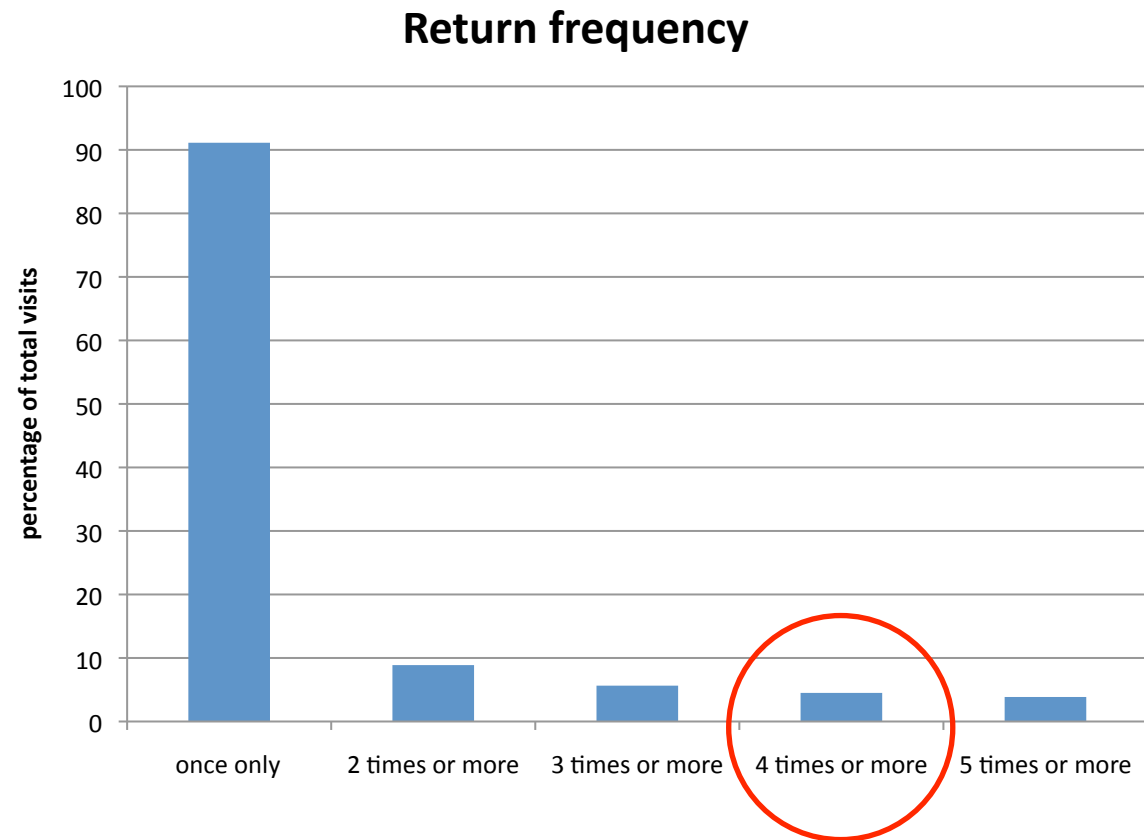
Graph Generated by SiteCatalyst using Report Accelerator at 9:46 AM PST, 12 Feb 2009

country	visits	percentage
China	51,285	26.80%
United States	28,916	15.10%
Taiwan	25,753	13.40%
Korea - South	12,934	6.80%
Hong Kong	8,826	4.60%
Brazil	4,648	2.40%
Canada	3,917	2.00%
Germany	3,077	1.60%
United Kingdom	2,928	1.50%
Japan	2,581	1.30%
Italy	2,579	1.30%
Spain	2,557	1.30%
Russian Federation	2,448	1.30%
Saudi Arabia	2,021	1.10%

Growth of the Database: Users and Sessions

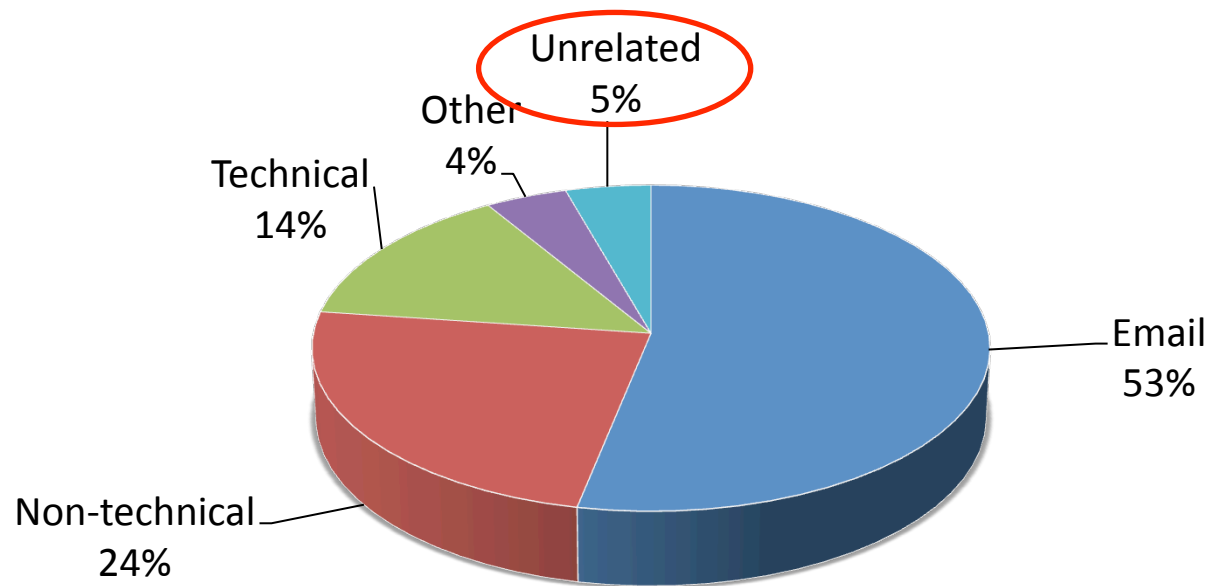


Repeat users



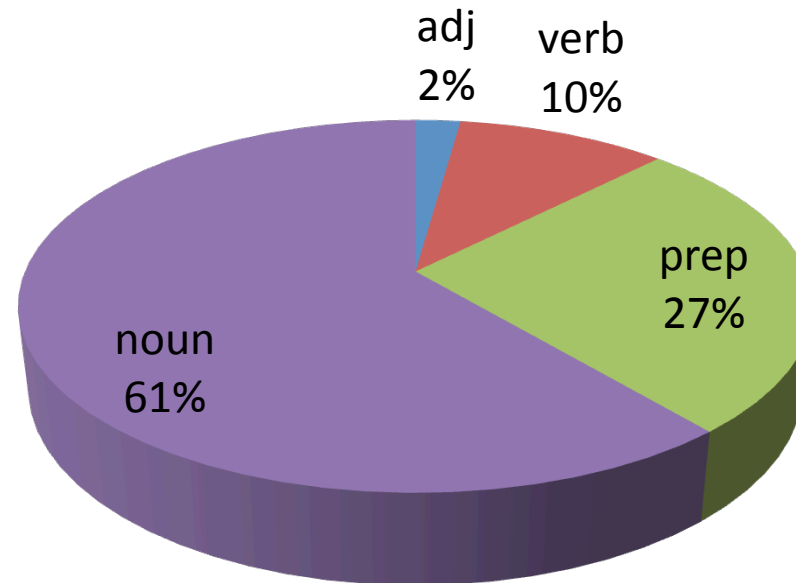
Collected Data

Writing Domains: By Number of Sentences



Frequent Users (2/10/09)

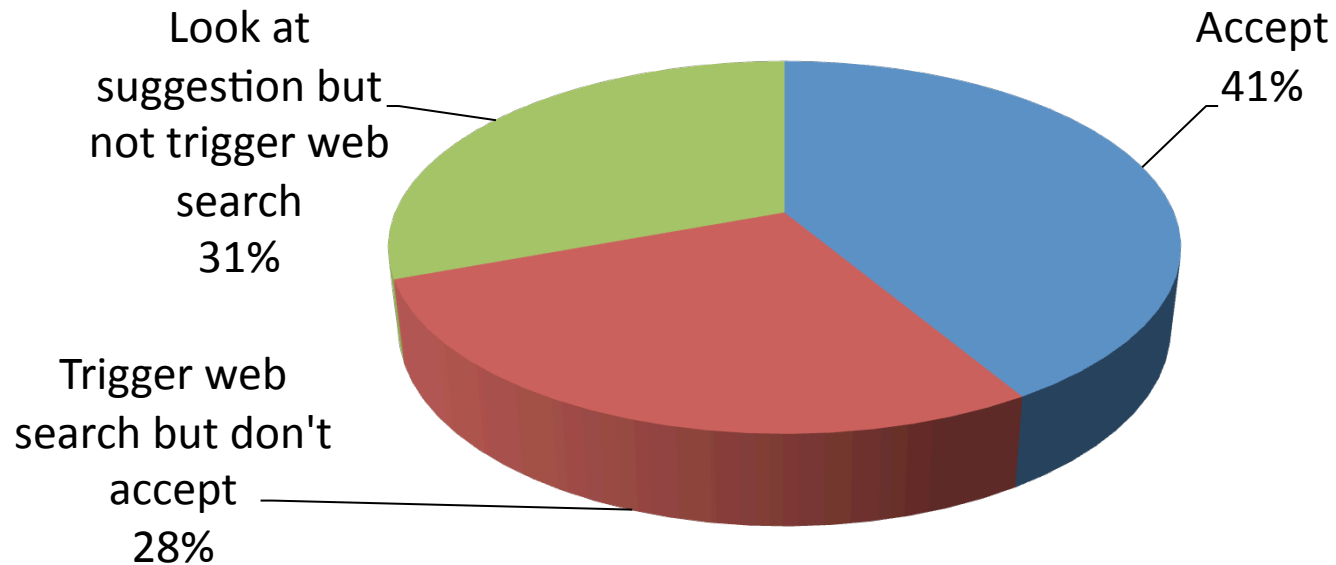
Frequent Users	578
Sessions	5,305
Session-Unique Sentences	39,944
Grammatical Error Flags	17,832



User interactions



Users Examine 87% of Suggestions



Conclusion: A significant number of users are inspecting the suggested rewrites and making a deliberate choice to accept it or not accept it.

Do users make the right choices?

To answer, need human evaluation:

- Time consuming, costly
- Inter-rater agreement (Tetreault & Chodorow)

BUT ... necessary for system development

- Single Annotator
- Internally consistent to measure relative performance during system development

To answer:

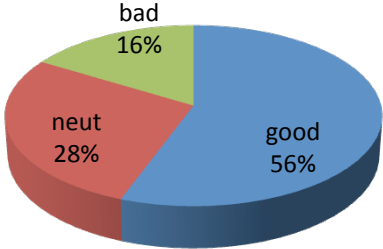
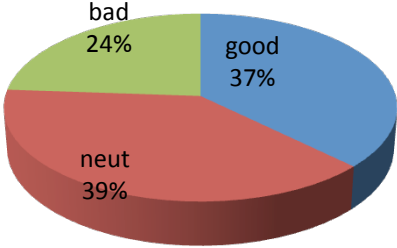
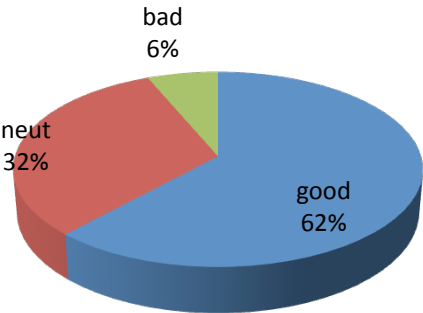
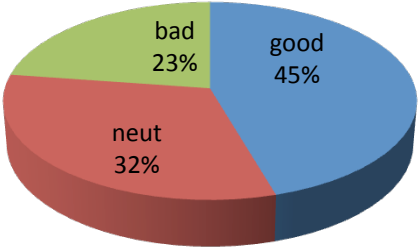
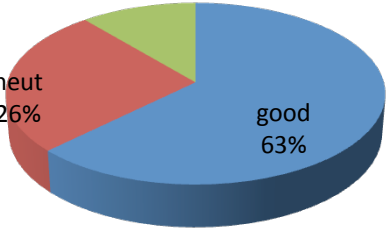
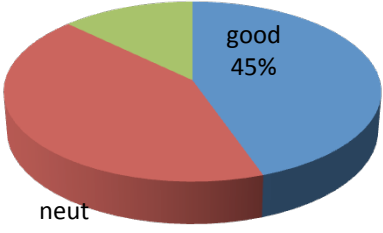
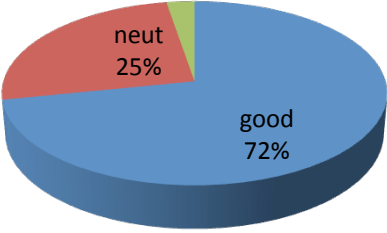
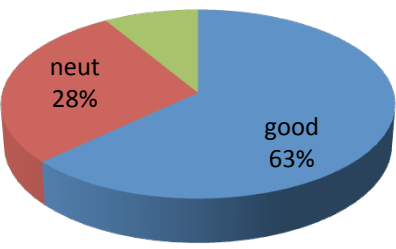
Do users make the right choices?

- Evaluated user data to date:
 - 34% of frequent user sessions: 6K flags
- From Evaluated Flags:
 1. Calculate performance for ALL suggestions.
 2. Calculate system performance for ONLY suggestions that were accepted.
 3. Compare ratios of good and bad flags.

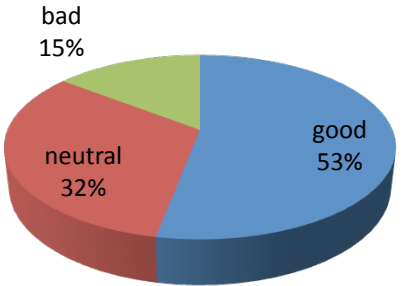
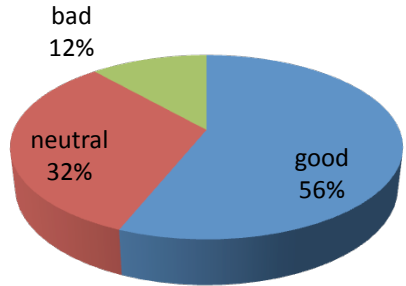
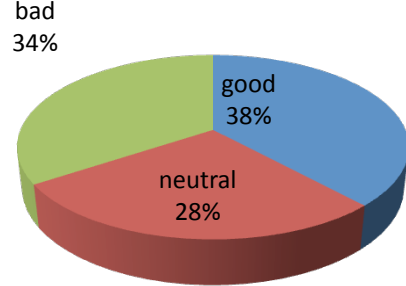
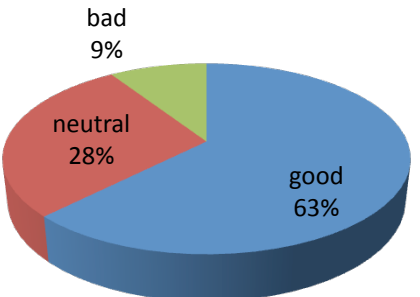
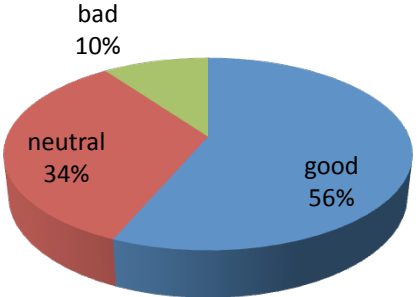
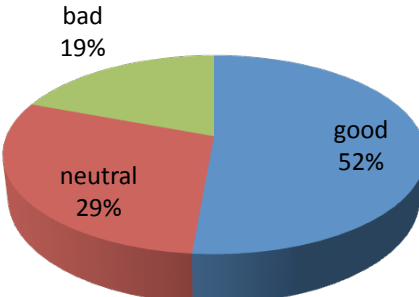
Evaluation Categories

Evaluation	SubEval	Description
Good	Correct Flag	The correction fixes a problem in the user input.
Neutral	Both Good	The suggestion is a legitimate alternative of a well-formed original input. <i>Ex: I like working/to work.</i>
	Misdiagnosis	The original input contained an error but the suggested rewrite neither improves nor further degrades the user input. <i>Ex: If you have fail machine on hand.</i>
	Both Wrong	An error type is correctly diagnosed but the suggested rewrite does not correct the problem. <i>Ex: "can you give me ^ suggestion" insert the instead of a</i>
	Non-ascii	A non-ascii or text processing mark-up character is in the immediate context. (Only applies to user data)
Bad	False Flag	The suggestion resulted in an error or would otherwise lead to a degradation over the original user input.

Error Type: Are users accepting the right suggestions?

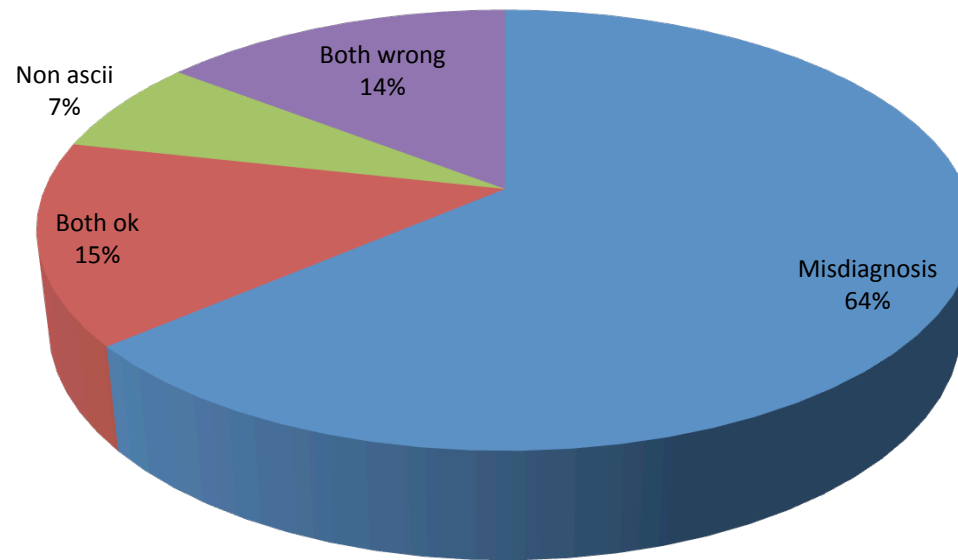
Noun-related	Prep-related	Verb-related	Adj-related																																
<p data-bbox="309 491 613 536">All Suggestions</p>  <table border="1" data-bbox="241 616 622 865"> <tr><th>Category</th><th>Percentage</th></tr> <tr><td>good</td><td>56%</td></tr> <tr><td>neut</td><td>28%</td></tr> <tr><td>bad</td><td>16%</td></tr> </table>	Category	Percentage	good	56%	neut	28%	bad	16%	<p data-bbox="761 491 1066 536">All Suggestions</p>  <table border="1" data-bbox="689 632 1084 880"> <tr><th>Category</th><th>Percentage</th></tr> <tr><td>good</td><td>37%</td></tr> <tr><td>neut</td><td>39%</td></tr> <tr><td>bad</td><td>24%</td></tr> </table>	Category	Percentage	good	37%	neut	39%	bad	24%	<p data-bbox="1205 491 1509 536">All Suggestions</p>  <table border="1" data-bbox="1142 561 1563 874"> <tr><th>Category</th><th>Percentage</th></tr> <tr><td>good</td><td>62%</td></tr> <tr><td>neut</td><td>32%</td></tr> <tr><td>bad</td><td>6%</td></tr> </table>	Category	Percentage	good	62%	neut	32%	bad	6%	<p data-bbox="1666 491 1971 536">All Suggestions</p>  <table border="1" data-bbox="1617 625 2033 874"> <tr><th>Category</th><th>Percentage</th></tr> <tr><td>good</td><td>45%</td></tr> <tr><td>neut</td><td>32%</td></tr> <tr><td>bad</td><td>23%</td></tr> </table>	Category	Percentage	good	45%	neut	32%	bad	23%
Category	Percentage																																		
good	56%																																		
neut	28%																																		
bad	16%																																		
Category	Percentage																																		
good	37%																																		
neut	39%																																		
bad	24%																																		
Category	Percentage																																		
good	62%																																		
neut	32%																																		
bad	6%																																		
Category	Percentage																																		
good	45%																																		
neut	32%																																		
bad	23%																																		
<p data-bbox="327 992 519 1037">Accepted</p>  <table border="1" data-bbox="224 1120 613 1353"> <tr><th>Category</th><th>Percentage</th></tr> <tr><td>good</td><td>63%</td></tr> <tr><td>neut</td><td>26%</td></tr> <tr><td>bad</td><td>11%</td></tr> </table>	Category	Percentage	good	63%	neut	26%	bad	11%	<p data-bbox="788 992 981 1037">Accepted</p>  <table border="1" data-bbox="689 1120 1070 1353"> <tr><th>Category</th><th>Percentage</th></tr> <tr><td>good</td><td>45%</td></tr> <tr><td>neut</td><td>42%</td></tr> <tr><td>bad</td><td>13%</td></tr> </table>	Category	Percentage	good	45%	neut	42%	bad	13%	<p data-bbox="1254 992 1447 1037">Accepted</p>  <table border="1" data-bbox="1160 1120 1545 1353"> <tr><th>Category</th><th>Percentage</th></tr> <tr><td>good</td><td>72%</td></tr> <tr><td>neut</td><td>25%</td></tr> <tr><td>bad</td><td>3%</td></tr> </table>	Category	Percentage	good	72%	neut	25%	bad	3%	<p data-bbox="1715 992 1908 1037">Accepted</p>  <table border="1" data-bbox="1617 1120 2011 1369"> <tr><th>Category</th><th>Percentage</th></tr> <tr><td>good</td><td>63%</td></tr> <tr><td>neut</td><td>28%</td></tr> <tr><td>bad</td><td>9%</td></tr> </table>	Category	Percentage	good	63%	neut	28%	bad	9%
Category	Percentage																																		
good	63%																																		
neut	26%																																		
bad	11%																																		
Category	Percentage																																		
good	45%																																		
neut	42%																																		
bad	13%																																		
Category	Percentage																																		
good	72%																																		
neut	25%																																		
bad	3%																																		
Category	Percentage																																		
good	63%																																		
neut	28%																																		
bad	9%																																		

Domains: Are users accepting the right suggestions?

Email	Non-technical	Technical																								
<p data-bbox="421 523 667 571">Suggestions</p>  <table border="1"> <tr><th>Category</th><th>Percentage</th></tr> <tr><td>good</td><td>53%</td></tr> <tr><td>neutral</td><td>32%</td></tr> <tr><td>bad</td><td>15%</td></tr> </table>	Category	Percentage	good	53%	neutral	32%	bad	15%	<p data-bbox="990 523 1236 571">Suggestions</p>  <table border="1"> <tr><th>Category</th><th>Percentage</th></tr> <tr><td>good</td><td>56%</td></tr> <tr><td>neutral</td><td>32%</td></tr> <tr><td>bad</td><td>12%</td></tr> </table>	Category	Percentage	good	56%	neutral	32%	bad	12%	<p data-bbox="1572 523 1818 571">Suggestions</p>  <table border="1"> <tr><th>Category</th><th>Percentage</th></tr> <tr><td>good</td><td>38%</td></tr> <tr><td>neutral</td><td>28%</td></tr> <tr><td>bad</td><td>34%</td></tr> </table>	Category	Percentage	good	38%	neutral	28%	bad	34%
Category	Percentage																									
good	53%																									
neutral	32%																									
bad	15%																									
Category	Percentage																									
good	56%																									
neutral	32%																									
bad	12%																									
Category	Percentage																									
good	38%																									
neutral	28%																									
bad	34%																									
<p data-bbox="452 976 645 1024">Accepted</p>  <table border="1"> <tr><th>Category</th><th>Percentage</th></tr> <tr><td>good</td><td>63%</td></tr> <tr><td>neutral</td><td>28%</td></tr> <tr><td>bad</td><td>9%</td></tr> </table>	Category	Percentage	good	63%	neutral	28%	bad	9%	<p data-bbox="1025 976 1218 1024">Accepted</p>  <table border="1"> <tr><th>Category</th><th>Percentage</th></tr> <tr><td>good</td><td>56%</td></tr> <tr><td>neutral</td><td>34%</td></tr> <tr><td>bad</td><td>10%</td></tr> </table>	Category	Percentage	good	56%	neutral	34%	bad	10%	<p data-bbox="1585 976 1778 1024">Accepted</p>  <table border="1"> <tr><th>Category</th><th>Percentage</th></tr> <tr><td>good</td><td>52%</td></tr> <tr><td>neutral</td><td>29%</td></tr> <tr><td>bad</td><td>19%</td></tr> </table>	Category	Percentage	good	52%	neutral	29%	bad	19%
Category	Percentage																									
good	63%																									
neutral	28%																									
bad	9%																									
Category	Percentage																									
good	56%																									
neutral	34%																									
bad	10%																									
Category	Percentage																									
good	52%																									
neutral	29%																									
bad	19%																									

What do users do with neutral flags?

Neutral Categories: “both wrong” and “misdiagnosis” 78% of neutral flags



Inspect >15.5K Flags to Accept 6.4K

Neutral Flags not accepted but sentence edited to produce no flag

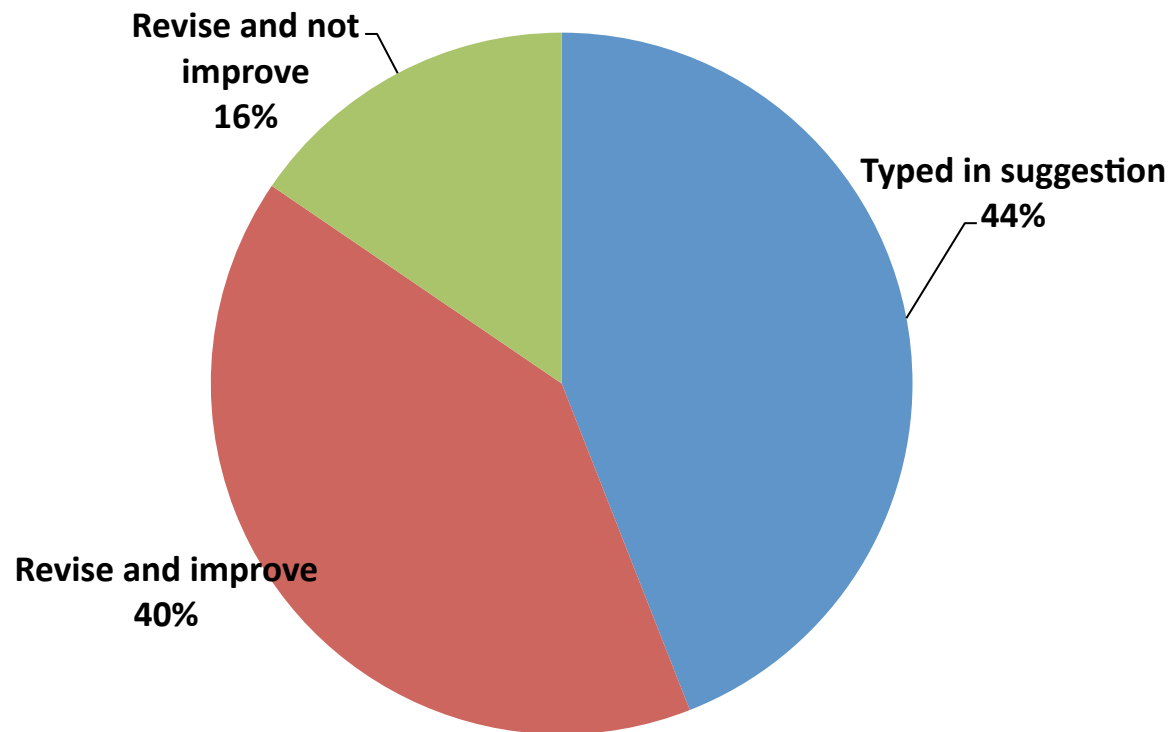
- I don't know that you knew or not , this early morning i got a from head office ...
 - suggestion: delete “from”
I don't know that you knew or not , this early morning I heard from the head office ...
- Please play with the software and Friday I will be by to work with any questions you may regarding it.
 - suggestion: regarding→regard
Please play with the software and Friday I will be by to work with any questions you may have regarding it.

From 1,349 sentences with neutral flags found 215 subsequently submitted “similar” strings with no error flag.

Users not accept suggestion but did something ELSE to make the flag go away.

Users improve 40% of the time

Not Accept Suggestion but Revise Sentence



Identifying the location of an error can help the user.

Conclusions

- Traffic: There is an interest in ESL proofing tools
- Even current state-of-the-art error correction can be useful for ELLs:
 - Users do not accept proposed corrections blindly – they are selective in their behavior
 - Users make informed choices – they can distinguish correct suggestions from incorrect ones
 - Sometimes just identifying the location of an error enables the users to repair the problem themselves

www.eslassistant.com