Microsoft Research

Butler Hill Group

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

Claudia Leacock, Butler Hill Group

Michael Gamon, Microsoft Research

Chris Brocket, 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

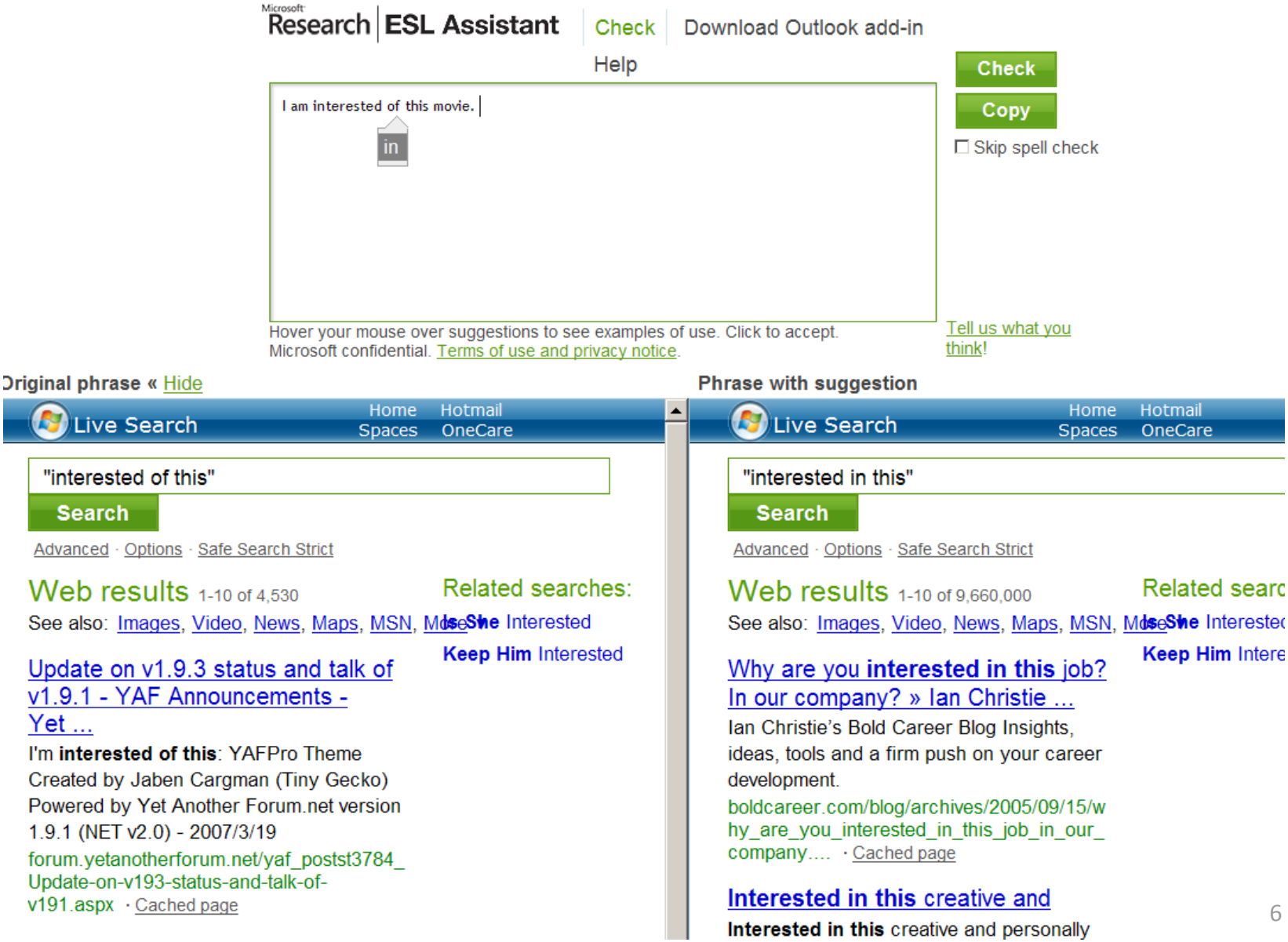
- On Saturday, I with my classmate went *eating/to eat.
- Hope you will *happy/be happy in Taiwan.
- I *teached/taught him all the things I know.

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

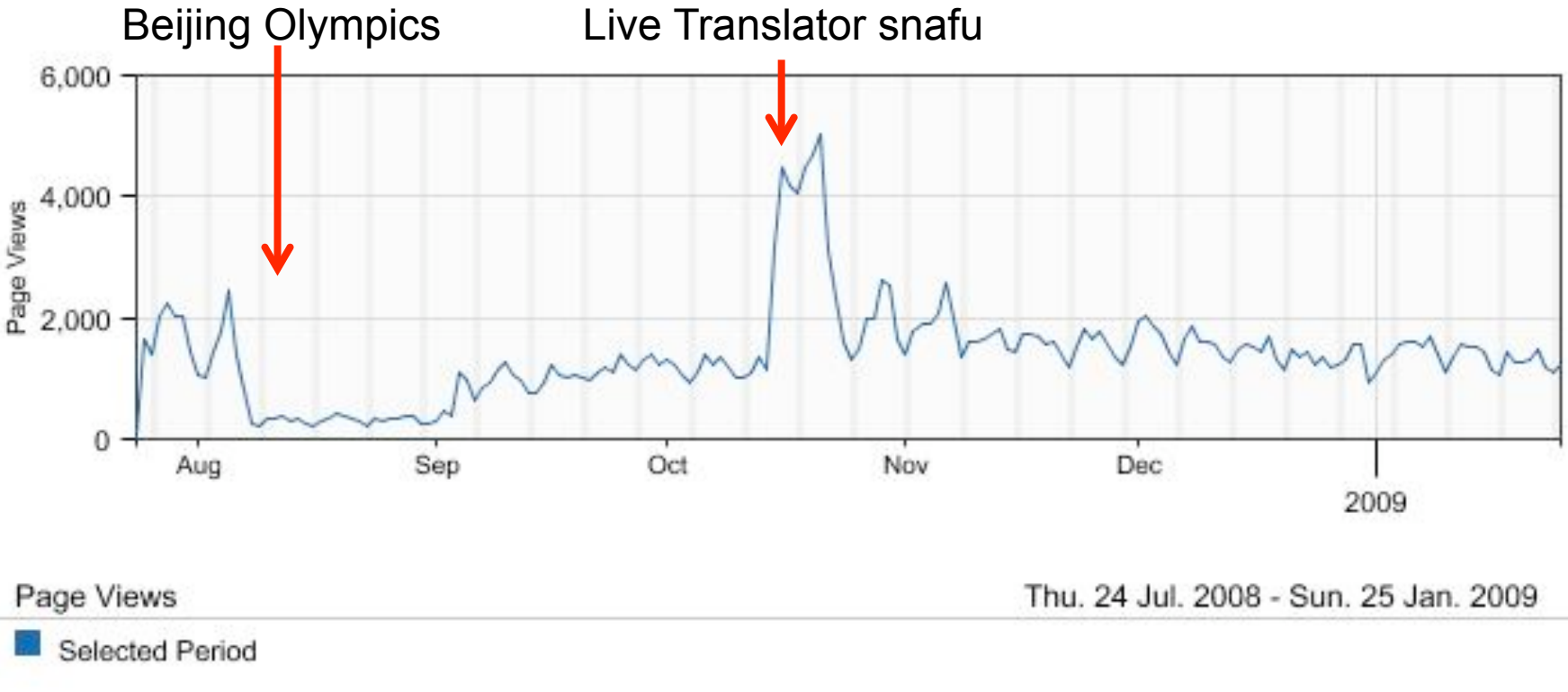
ESL Assistant User Interface



Page Views per Day

Traffic via website links:

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

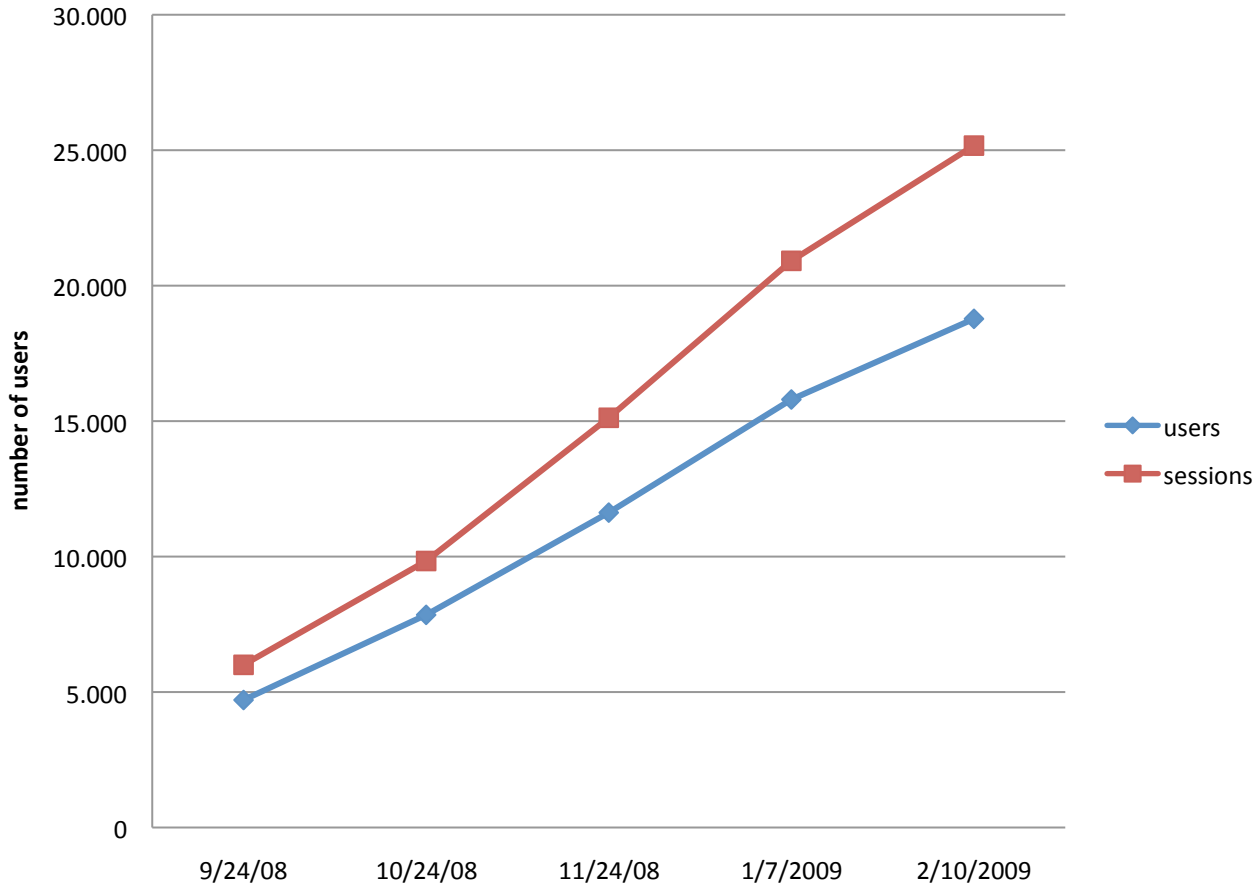


User Location

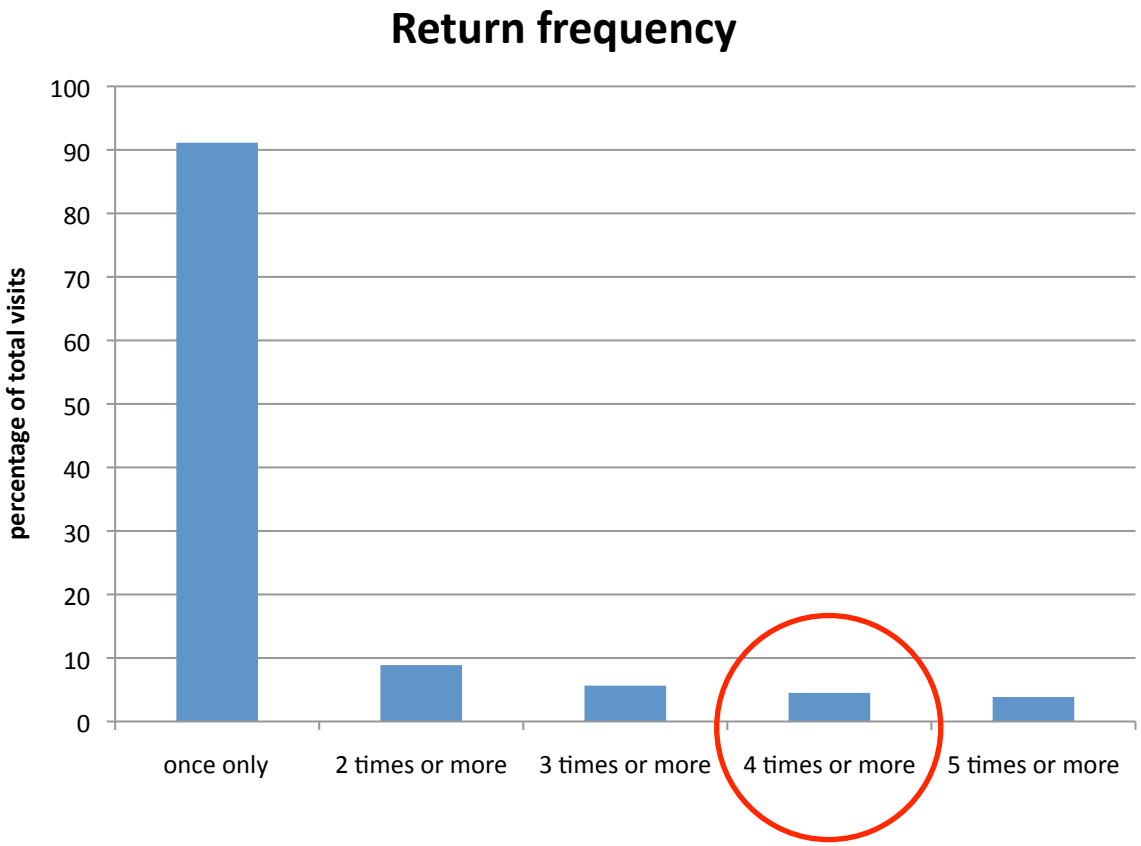
Graph Generated by SiteCatalyst using Report Accelerator at 9:46 AM PST, 12 Feb 2009

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%

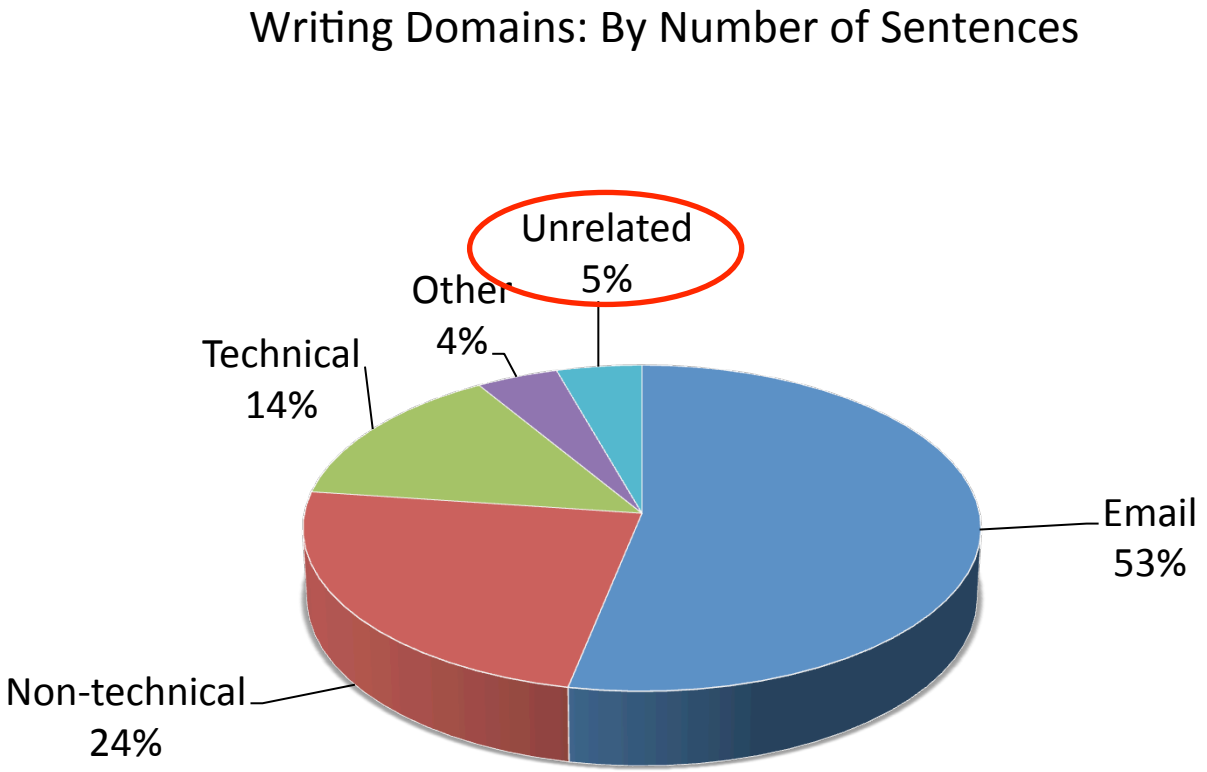
Growth of the Database: Users and Sessions



Repeat users

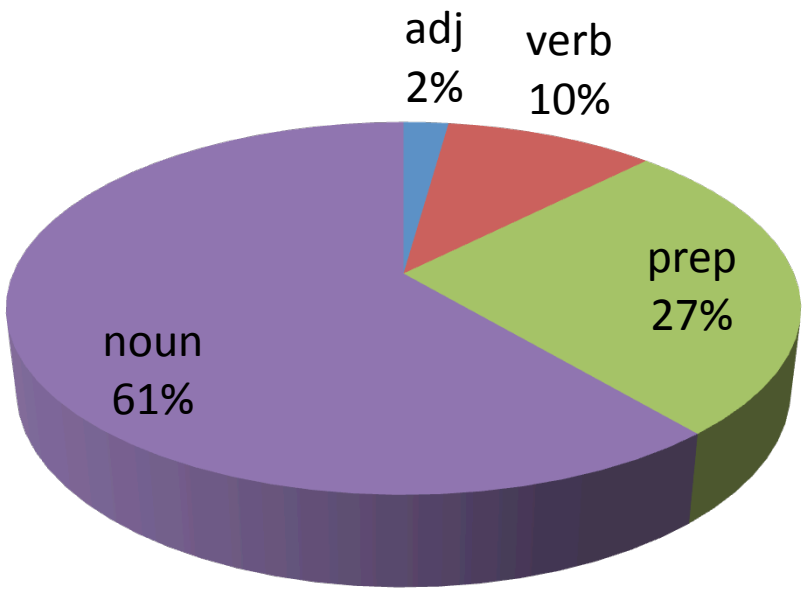


Collected Data



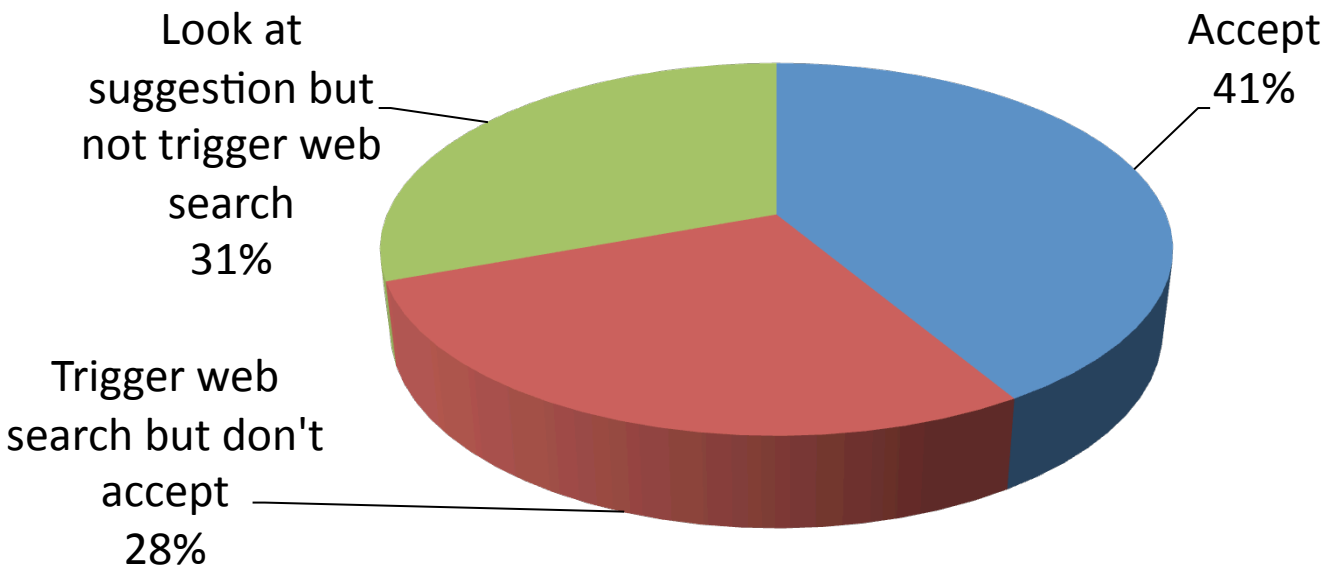
Frequent Users (2/10/09)

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

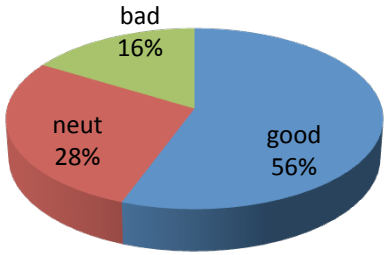
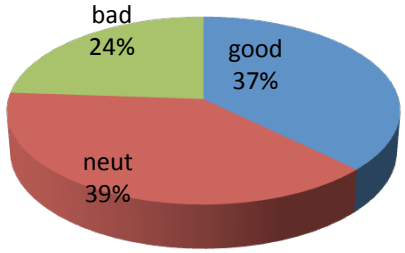
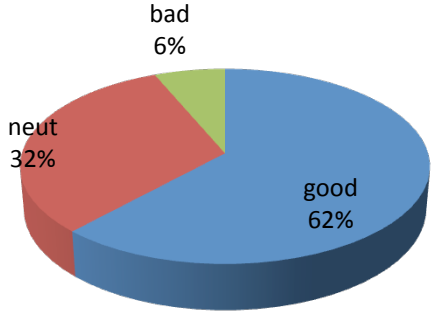
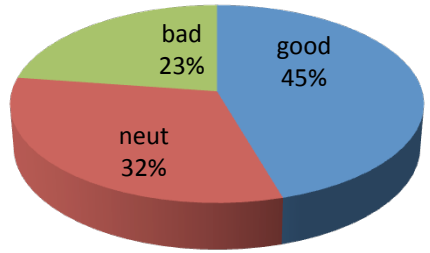
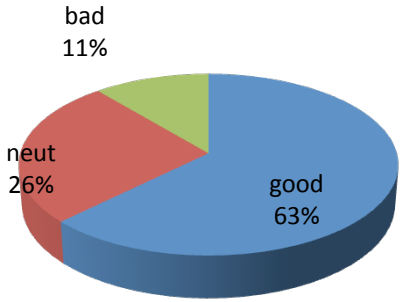
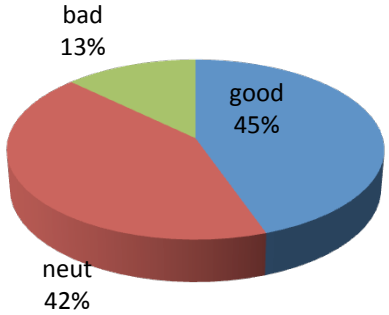
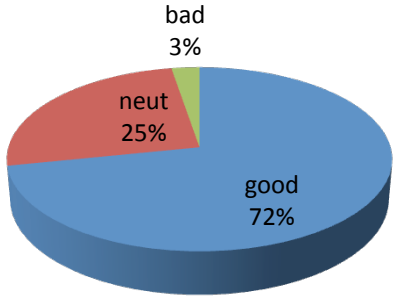
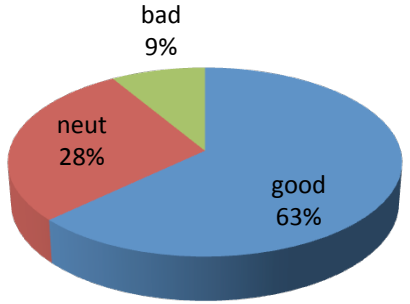
- From Evaluated Flags:

  1. Calculate performance for ALL suggestions.
  2. Calculate system performance for ONLY suggestions that were accepted.
  3. Compare ratios of good and bad flags.

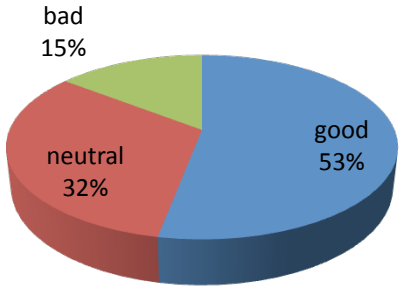
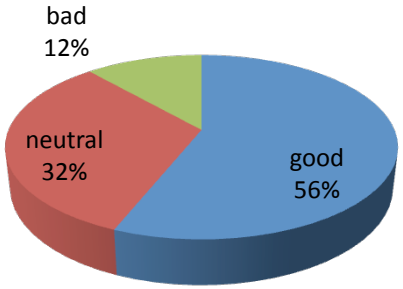
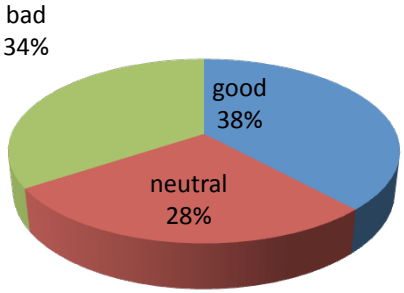
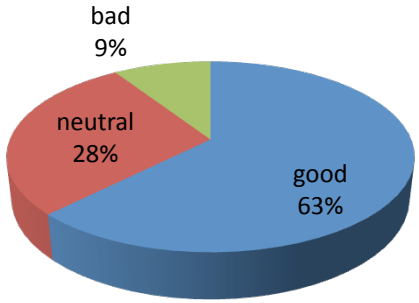
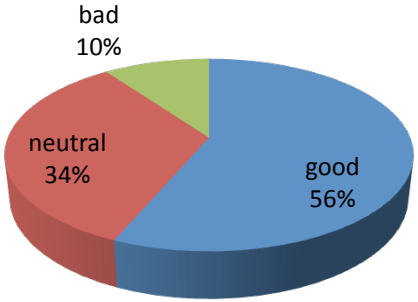
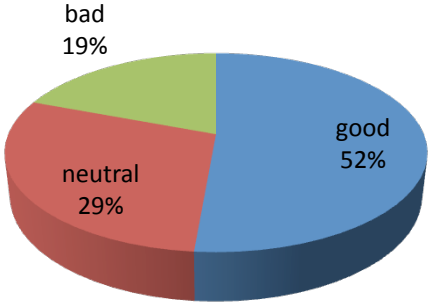
Evaluation Categories

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. *Ex*: I like working/to work.
	Misdiagnosis	The original input contained an error but the suggested rewrite neither improves nor further degrades the user input. *Ex:* If you ~~have~~ fail machine on hand.
	Both Wrong	An error type is correctly diagnosed but the suggested rewrite does not correct the problem. Ex: "can you give me ^ suggestion" insert the instead of a
	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?

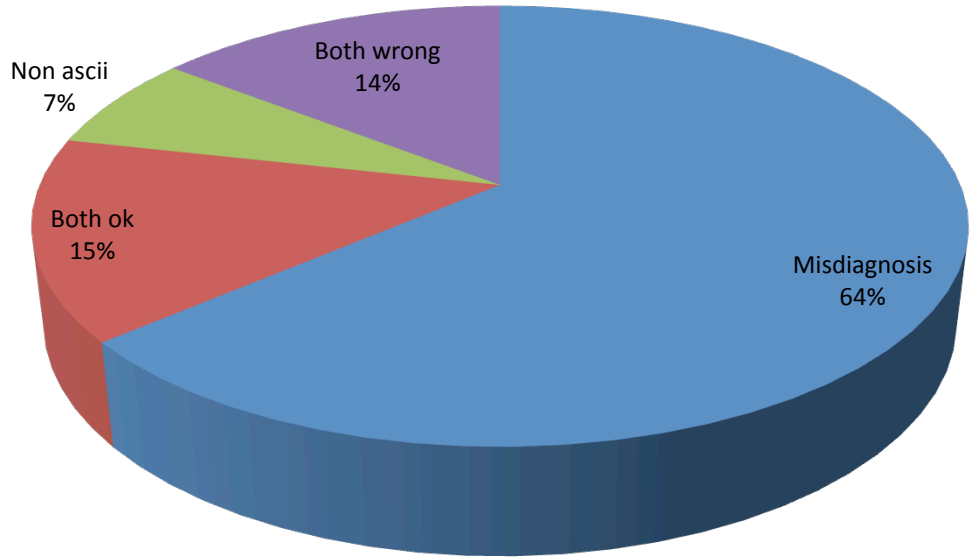
Noun-related	Prep-related	Verb-related	Adj-related
All Suggestions	**All Suggestions**	**All Suggestions**	**All Suggestions**
			
Accepted	**Accepted**	**Accepted**	**Accepted**
			

Domains: Are users accepting the right suggestions?

Email	Non-technical	Technical
Suggestions 	**Suggestions** 	**Suggestions** 
Accepted 	**Accepted** 	**Accepted** 

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

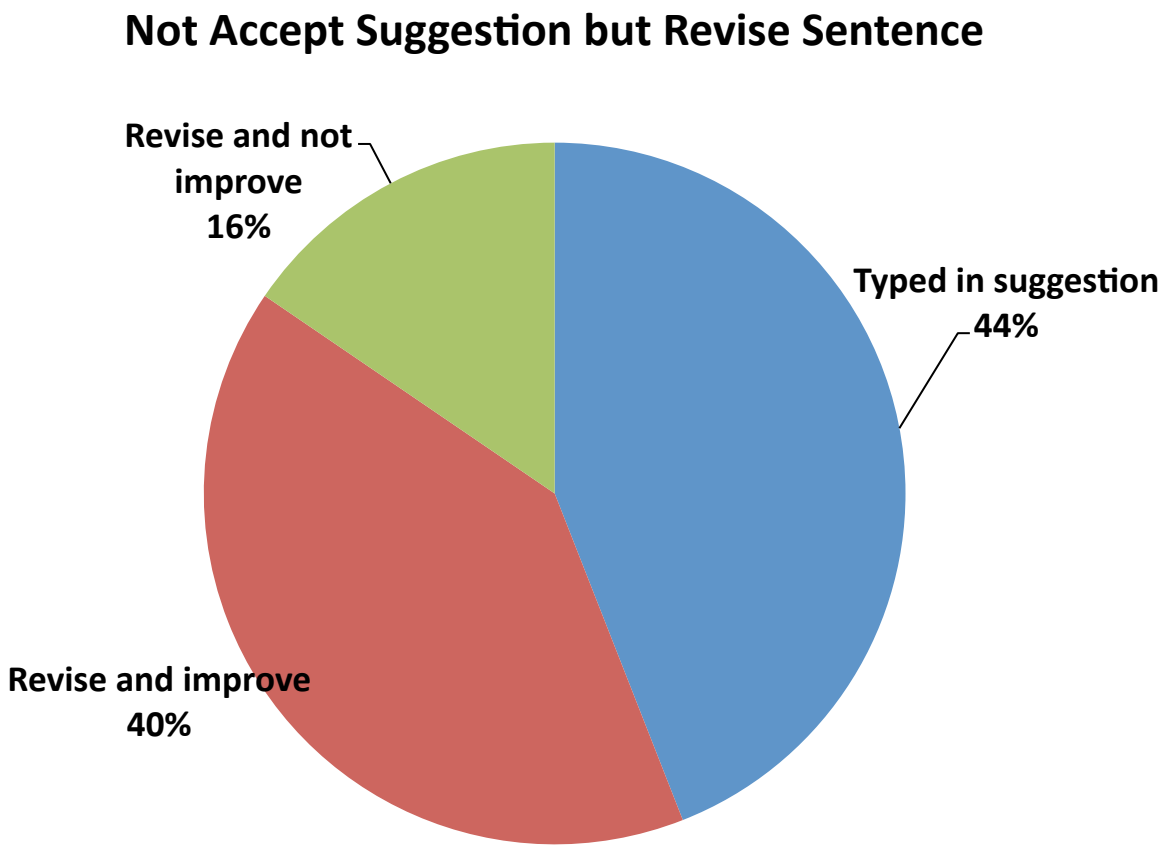
- I don't know that you knew or not , this early morning i got a from head office ...
  - suggestion: delete “from”

  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



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