

Google as a Corpus Tool?

In October of last year there was an interesting series of postings on the ETJ discussion list which forms the basis of this article. The Internet is a powerful tool, but with any tool, there is the potential to use it in unintended and perhaps inadvisable ways. Is using Google as a guide to English usage one of them? Read this and then decide for yourself.

It started out with one member asking about the correctness of "I'd appreciate if" deleting the object "it."

The "standard" corpora (see references) had no entries for "appreciate if" and only a handful for "appreciate it if" -- not very good evidence for

anything. Then one member looked on Google and got tons of hits -- over 90,000, with 59,000 for "I hate when" another phrase under discussion.

So, what can we conclude from this? We can certainly say that these forms are used -- but by whom, and in what context? Does the fact that they are "used" mean that we should be teaching them?

Edward Estry pointed out that even searching for a clearly incorrect usage such as "I would of gone" resulted in 344 hits. I did a similar search on "would of been" and found close to 1000 examples, including this lovely one: "When I

turned around and found Cal standing right next to me, he would of been 3 center meters away, I was in total shock."

So, where does this leave us? Estry observes, "Of course language is changing all the time, but where do we as teachers draw the line? Would you say to your students, 'Well, some native speakers use this structure incorrectly so it might be OK.???'"

Going back to the original query concerning "appreciate if," even if we try to reduce the hits to sites where one would expect "educated" usage, we still find numerous instances. (The figures after the "/" are the number of hits with "it" properly inserted in the phrase.)

"appreciate if" ".ac.uk" --> 5000+ hits /7200+ hits

"appreciate if" ".edu" --> 20000+ hits /46000+ hits

Thus it would seem that there are quite a few educated users, and from the ratio of deleted-to-included, that the form is more prevalent in the UK than in the US. (The addition of the domain name does not perfectly restrict the search to that domain, but rather to any sites that have the stipulated phrase anywhere on their page.)

Charles Kowalski observed, "Are we seeing a gradually increasing flexibility in de facto rules of usage, with the result that the empty "it" in the object slot can be left unstated and assumed to be there (like the "that" introducing an embedded sentence)? Or shall we go with the prescriptive view that these expressions are just plain wrong?"

Thus Google might be useful for capturing data on a language change in progress, but still, it doesn't tell us anything about whether the new form is generally accepted.

Furthermore, it doesn't situate this change in a larger context. What class of verbs are undergoing this change? What genres of English does

The screenshot shows a Google search interface with the query "appreciate if" ".edu" and the search button "Google 検索". Below the search bar, there are radio buttons for "ウェブ全体から検索" (selected) and "日本語のページを検索". The search results are displayed in a list format, showing various snippets of text from different websites, many of which include the phrase "I would appreciate if" or "I appreciate if".

検索オプション 表示設定 言語ツール ヘルプ

Google "appreciate if" ".edu" Google 検索

ウェブ全体から検索 日本語のページを検索

ウェブ イメージ グループ ディレクトリ

全言語のページから "appreciate if" ".edu" を検索しました。

[Will Muslims link to rebuttals of Dr. Bucaille?](#)
... [edu/~kaldirog/islam/articles/bucaille.html](#) http://multimedia.ecn.purdue.edu/~kaldirog/islam/articles
I would **appreciate if** you could link ...
[answering-islam.org/Campbell/linking.html](#) - 6k - [キャッシュ](#) - [関連ページ](#)

[header.jpg](#)
...). Would **appreciate if** you can help us send this out through your society ... Return-Path:
<maelee@edb.gov.sg> Received: from mx1.andrew.cmu.edu (MX1.andrew.cmu.edu ...
[ssa.web.cmu.edu/notices-edb.htm](#) - 20k - [キャッシュ](#) - [関連ページ](#)

[Re: Agency + Civil Procedure questions](#)
... [topeka.wuacc.edu](#)> Sent: Monday, March 20, 2000 11:05 AM Subject: Agency + Civil Procedure
questions > Dear all, > > I would very much **appreciate if** subscribers ...
[legalminds.lp.findlaw.com/list/canadalawyers/msg00046.html](#) - 9k - [キャッシュ](#) - [関連ページ](#)

[Agentcities projects at RMIT](#)
... We also greatly **appreciate if** you could send us email to let us know why and how ... Wednesday,
04-Dec-2002 20:42:46 EST Email: [WebManager@cs.rmit.edu.au](#) Copyright ...
[www.cs.rmit.edu.au/agents/protocols/](#) - 9k - 2003年3月23日 - [キャッシュ](#) - [関連ページ](#)

[Matt Schuette II Email](#)
... form, noting that all fields are optional (I would **appreciate if** you filled ... To: Email
([schuette@umr.edu](#)). ...
[www.umr.edu/~schuette/mailform.html](#) - 6k - [キャッシュ](#) - [関連ページ](#)

[Christopher Faylor - \[cobbler@stanford.edu: Re: Serial blocking ...](#)
... 2000 14:39:16 -0800 (PST) Reply-To: [dmorris@alumni.brown.edu](#) In-Reply ... I would **appreciate**
if someone could actually > verify that my patch does the right thing. ...
[www.cygwin.com/ml/cygwin-developers/2000-12/msg00079.html](#) - 10k - [キャッシュ](#) - [関連ページ](#)

[www.math.niu.edu/~rusin/known-math/99/multigrd](#)
... > Now I am definitively going to library... > > I **appreciate if** you help me again ... Jun
Zhang * E-mail: [jzhang@cs.uky.edu](#) * * Department of Computer Science * URL ...
6k - [キャッシュ](#) - [関連ページ](#)

[FEMISA: feb97 : B95: Cuestionario \(fwd\)](#)
... [juleiac@leland.Stanford.EDU](#)> To: [womens-coalition@lists.Stanford.EDU](#) Subject: Cuestionario
(fwd ... would like to answer the survey, we would **appreciate if** you gave ...
[csf.colorado.edu/forums/femisa/feb97/0035.html](#) - 6k - [キャッシュ](#) - [関連ページ](#)

Re: [manet] delay in sending the next packet in DSR

this change apply to?

As pointed out by Michael Rundell, the Internet “is not a corpus at all according to any of the standard definitions: what it is is a huge ragbag of digital text, whose content and balance are largely unknown. It is, in the jargon, a highly “skewed” archive, in that some text-types are very well represented, and others are hardly present at all.”

But even having said all this, there are times when Google can provide a quick “reality” check. As a native English speaker who has been removed from the NS environment for quite a while, sometimes it is useful to confirm that there are other people out there who use a particular form, be it a vocabulary item, idiom, or syntactic phrase.

If you don't have a dictionary at hand, Google can often tell you if you are spelling a word correctly although you will find plenty of pages with spellings. Google often will come up with, “Are you sure you didn't mean xxxxxx?” when you input a wrong form.

Nevertheless, if you want to find examples for use in class, Google is not ideal:

1. You can only search for specific words, not word categories or inflected forms.

2. There is no control over the educational level, nationality, or other

characteristics of the creators of the utterances found.

3. The results are not in an easy-to-read format.

On the positive side,

1. It is much more accessible than any corpus.

2. The database is huge compared to any existing corpus.

3. The index sites include blogs and discussions which come very close to spoken language whereas much of the data in formal corpora are from more formal written styles.

If, however, you are looking for are more examples of a specific grammatical pattern to present to your students, perhaps Charles Kelly's corpus of sentences from the Voice of America's Special English program would be useful. This avoids the problem of the sentences retrieved being too difficult for most students to understand, particularly out of context:

<http://www.manythings.org/voa/sentences/htm>

And now, the good news has been saved for last. If you do insist on using the Internet as a corpus, Bill Pellowe reports that an offshoot of University of Liverpool, Webcorp, has developed an online program that can perform a search with “wild-cards” and numerous other specifications, and produce the results in a readily readable form.

Caveat Clicker!

References

Rundell, Michael (2000). “The biggest corpus of all”, *Humanising Language Teaching*, Year 2; Issue 3; May 2000
Available:
<http://www.hltmag.co.uk/may00/idea.htm>

Webcorp

<http://www.webcorp.org.uk/index.html>

Online Corpora

BNC (British National Corpus)
<http://sara.natcorp.ox.ac.uk/lookup.html>

Cobuild

<http://titania.cobuild.collins.co.uk/form.html#democonc>

Michigan Corpus of Academic Spoken English

<http://www.hti.umich.edu/m/micase/>

Virtual Language Center

<http://www.edict.com.hk/concordance/>

Information on concordancing

Michael Barlow's site “Corpus Linguistics”
<http://www.ruf.rice.edu/~barlow/corpus.html>

Teaching English in Japan - Corpora directory,

<http://www.teaching-english-in-japan.net/directory/cat/85>

Teaching English in Japan - Text analysis software,

<http://www.teaching-english-in-japan.net/directory/cat/86>

Tim Johns' Data Driven Learning page

<http://web.bham.ac.uk/johnstf/timconc.htm>

Ask the Techie

This column is for you to ask your own questions about using computers and the Internet with your students. Send your questions to the editor, Tom Robb at tom@robb.net

Q: I want to evaluate some readings to find predict how difficult they might be for my students. Is there any software that can help?

Your first resource should probably be MS Word, which can report the Flesch “reading ease” (higher is easier) and grade level of the passage in question. In order to use this function, however, you first need to access your preferences, and under the proofing preferences, check the

box for “readability.” Once that is done, MS Word will report the readability each time you finish spell checking your document.

Another useful tool is Paul Nation's “Vocabulary Profiler” (http://www.er.uqam.ca/nobel/r21270/cgi-bin/webfreqs/web_vp.html). You paste in your text, click “Submit” and it comes up with an analysis of your text with the words color-coded for their vocabulary level.

Counts	
Words	1151
Characters	5904
Paragraphs	58
Sentences	51
Averages	
Sentences per Paragraph	2.0
Words per Sentence	19.3
Characters per Word	4.6
Readability	
Passive Sentences	5%
Flesch Reading Ease	59.2
Flesch-Kincaid Grade Level	9.7