GLOTTODIDACTICA XXXIII (2007)
ADAM MICKIEWICZ UNIVERSITY PRESS POZNAŃ

PAWEŁ SCHEFFLER
Uniwersytet im. Adama Mickiewicza w Poznaniu

When Intuition Fails us: the World Wide Web as a Corpus

ABSTRACT. In some respects corpus linguistics has made a significant contribution to foreign language (L2) instruction: for example, reference tools like dictionaries and grammar books are at present enriched by various types of information derived from corpora. However, as far as teachers' and students' use of corpora is concerned, the impact of corpus linguistics has been rather limited. This article demonstrates how teachers and learners of English as a foreign language can use the World Wide Web as a corpus. More specifically, it shows how the WWW can help teachers and learners in making L2 acceptability judgements.

1. INDETERMINATE INTUITIONS

There is one question that both non-native teachers and learners of English inevitably ask themselves at some point: "Is this particular phrase something that a native speaker would write or say?" Teachers may ask this question when marking homework writing assignments. Learners may ask this question when doing their homework writing assignments. Before I suggest one quick way of answering it, I would like to say a few words about why such questions are asked in the first place, and what it is that makes answering them difficult, even for very proficient users of English.

It is a well-known fact in SLA research that even advanced learners of English do not have reliable intuitions concerning the target language. For example, Schachter, Tyson and Diffley (1976) demonstrate that high-intermediate and advanced learners of English cannot make clear grammaticality judgements of sentences containing certain types of relative clauses. They propose that whereas for native speakers strings of English words can be divided into grammatical/acceptable and ungrammatical/unacceptable, for non-native speakers many strings are indeterminate:

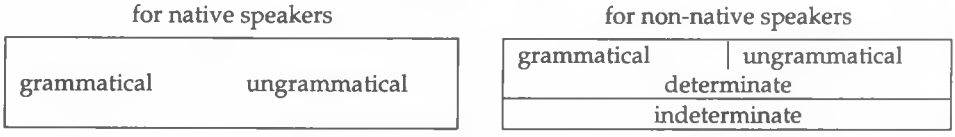

Figure 1: Set of strings of English words (adapted from Schachter, Tyson and Diffley 1976: 70)

Bley-Vroman (1989) suggests that this indeterminacy results from the fact that the rule system underlying learners' judgements is incomplete. But the problems that learners have with judging and producing English phrases and sentences are not restricted to the syntactic rules of the target language. There is also the question of native-like selection of lexical items. There is now strong empirical evidence (e.g. Sinclair 1991) that speech or text production is to a large degree based on lexical frames. This means that when a native speaker selects a particular word, then this selection influences certain other grammatical and lexical choices that are made. Some of these choices are fairly specific: for example, the phrase *set eyes on* attracts a pronominal subject, the auxiliary *has*, and either the adverb *never* or a conjunction like *the moment / the first time*. However, others are much less concrete, as is the case with the verb *happen*, which is normally used to talk about unpleasant things (e.g. accidents).

This dualistic view of language, i.e. language as a rule-based system and language as a 'frame-based system' implies that a learner of English, especially an advanced one, may have a native-like mastery of the former without having a comparable mastery of the latter. This is in fact what Pawley and Syder (1983) state often happens: learners are capable of producing speech that is grammatical and fluent, but which is not characterised by native-like selection of lexical frames.

There are three avenues which learners and teachers can explore when their linguistic knowledge and intuition fail them. The first is to consult a native speaker. The main problem with this solution is that one does not normally have a native speaker waiting to be consulted whenever one is engaged in producing or correcting a piece of writing. The second option is to look up a suspect phrase in a grammar book or a dictionary. Grammar books and dictionaries are more accessible to learners and teachers than native speakers are, and they do not mind being consulted over and over again, but they cannot always give information on just the problem that one is grappling with. This leaves them with the third option, namely consulting a corpus, and it is this option that I would like to recommend.

Roughly speaking, a corpus is a (large) collection of electronic texts in a particular language. There are a number of corpora of the English language (for example the British National Corpus) which have been carefully compiled in order to reflect different varieties of the language. These corpora are

used by linguists and grammarians in general descriptions of English and in the process of compiling dictionaries and writing grammar books, but they can also be used by learners and teachers working with the language.

The corpus that I would like to recommend to learners and teachers of English is the World Wide Web. Of course, it differs from corpora like the BNC in not being designed for use by linguists, teachers or learners. However, it contains a huge amount of authentic linguistic data, and it is available to anyone with access to the Internet. I will illustrate how the WWW can be used by teachers of English in marking student compositions in section 4.

2. LINGUISTIC CORPORA, THE INTERNET, AND FOREIGN LANGUAGE LEARNING AND TEACHING

In some areas, linguistic corpora have made a significant contribution to the process of foreign language instruction. This contribution is most visible in the area of reference tools, that is, dictionaries and grammar books (Meunier 2002). For example, dictionaries now provide frequency and register information, and grammar books often use authentic examples to illustrate particular points being discussed. In other areas, however, the impact of corpora and corpus research has been rather limited. One methodological procedure which appears to have gained some popularity is classroom concordancing. This enables learners to examine lists of corpus-based, computer-generated examples which show how a particular word or phrase behaves in its linguistic context. Through the examination of concordance lines, an inductive, data-driven approach to learning is encouraged (Johns 1994).

The value of the Internet and the WWW as computer-assisted language learning (CALL) applications has been recognised for some time now. Beatty (2003) discusses two main ways in which the Internet and the WWW can assist foreign language instruction: computer-mediated communication and the use of WWW resources. As far as the latter is concerned, learners and teachers can either use materials that have been especially designed for teaching purposes, or they can adapt those that have not been intended for language learning.

The idea that the WWW could be used as a corpus which learners and teachers consult to verify their knowledge or intuition does not seem to have attracted much attention so far. For example, Dudeney (2000: 22) admits that using the Web for linguistic searches of this kind is "one of the best tips to teach your students", but actually devotes very little space to it, and does not entertain the possibility that (non-native) teachers could also benefit

from this procedure. For Teeler and Gray (2000: 43) the usefulness of 'string searches' is restricted to "finding poetry, literary quotations, song lyrics and proper names." Finally, Hunston (2002: 170) discusses pedagogical uses of general corpora in the context of data-driven learning. She talks about students using concordance lines or sentences to answer questions like " 'Is it better to say x or y' " or " 'What is the difference between saying x and saying y?' ". In my view a more fundamental question that learners are faced with is 'Is it possible/natural to say x?'. To answer this question students do not need to analyse long sets of concordance lines, which for many may be a truly daunting task. All they need to do is to determine whether the problematic expression occurs in the data produced by native users of English. This is a much more modest goal than an analysis of concordance lines, but one which I believe to be realistic for the average student.

The relative absence of attention in the literature to the Web as a corpus which can be used to verify one's knowledge and intuition might explain why learners and teachers in general do not exploit the potential of the Web in this area. A survey which demonstrates that this potential is indeed unexplored by Polish teachers and learners of English is the subject of the next section.

3. THE WEB AS A CORPUS – A SURVEY

3.1. Data collection

To determine whether Polish teachers and learners of English use the Web as a corpus, a survey was conducted in which twenty teachers and one hundred advanced learners of English were asked the following question:

> In what way (if any) do you use the internet in your teaching/learning of English?

All the teachers were interviewed individually and notes were made of their answers. The students were approached in groups of about fifteen at a time during their regular class time. They were asked to write down their answers on paper and were given unlimited time to do so. They were not provided with any hints as to how the Internet can be used in teaching or in learning.

Both groups were randomly selected from teachers and students at the Poznań College of Modern Languages. The selection of this particular school was motivated by the fact that it is a tertiary level institution, a teacher training college, whose staff and students were in my opinion likely to make use of technology in the teaching/learning process. Also, the school is equipped with free-access, Internet-connected, computer labs, which makes access to web resources very easy.

3.2. Results and discussion

Out of the 100 students that were questioned, 13 admitted to not using the Internet at all in their learning. The raw figures for all the Internet resources referred to by the other student respondents are given in Table 1 below (the label 'web as a corpus' is my own; students usually wrote about 'searching for words or expressions on the Web'.)

Table 1

Internet resource	Number of students using the resource
1. dictionaries	58
2. news and magazines	35
3. EFL pages	19
4. literary materials	14
5. web as a corpus	10
6. radio	7
7. encyclopaedias	5
8. books	3
9. song lyrics	3
10. General corpora (BNC)	1

The results indicate that only ten per cent of the students surveyed do use the Web as a corpus, i.e. they use it to search for or to check the use of various linguistic items. As far as the teachers are concerned, the percentage of those who carry our linguistic searches of the Web was higher: six out of the twenty teachers that were questioned admitted to it. All the teachers in the sample except one claimed to be proficient Internet users and said they utilised it as a source of various supplementary materials for their classes. Two teachers reported using the British National Corpus as a means of verifying their linguistic intuitions. They did not, however, use in the classroom any corpus-based teaching procedures like concordancing.

4. SEARCHING THE WORLD WIDE WEB

To illustrate how the Web[1] can help teachers in judging the acceptability of student writing I chose ten problematic phrases from essays written by

[1] To carry out simple linguistic searches of the Web I recommend using one of the generally accessible search engines, for example Google or AltaVista. All one needs to do to carry out the search in this case is to enclose a given string in quotation marks. Another option would be to use computer programs available on the Internet which have been designed specifically for linguistic searches of the WWW (e.g.: http://www.webcorp.org.uk/wcadvanced.html). These linguistically-oriented programs are, however, slower and more difficult to use than general search engines.

Polish university students of English.[2] My choices were intuitive: as most teachers of English will admit, their intuition often tells them that a given phrase or construction may be problematic. Similarly, learners are often uncertain about the appropriateness of, say, particular collocations they have used, for instance simply because they have never encountered them before. It is questions such as these that searching the Web can help to answer.

In each case that is presented below, my intuitions were confirmed by the WWW Google search: no hits of the problematic phrases were displayed (except one example in which only one match was displayed). This, however, does not mean that I am always right: as the title of this article says, intuition often fails us and the corpus quite often proves me wrong.

In addition to helping teachers and learners identify problematic phrases, their intuition (or their knowledge) often suggests alternative words or expressions. These can also be verified with the help of the Web. In the examples below I first show the sentences with the problematic phrases, and then demonstrate how the WWW corpus can help us find fully acceptable equivalents.

1. *Sometimes it occurs to be controversial when, for example, a black mother (using donor eggs) gives birth to a white child.*

The problem in this case concerns the choice of the verb. If *occurs* is replaced with *turns out* 14 hits are displayed, the first of them being the following:

> Rhetoric for Engineers: Netiquette
>
> If you bring up a subject and **it turns out to be controversial**, lurk until you can participate without losing your temper. If you can keep your temper, ... www.tcnj.edu/~rgraham/rhetoric/netiquette.html - 7k -

2. *In small traditional shops one cannot come across such improbably good bargains as at Hit or Leclerc, but one can at least enjoy his independence.*

What I did not like in the highlighted phrase was the choice of the adverb. When I replaced it with *incredibly* I found 11 matches, for example:

> Charthouse Data Management Ltd - Why Charthouse?
>
> Aside from recieving (sic) extortionate quotes, you may also get what would seem to be **incredibly good bargains**. But are you really saving yourself any money if ... www.charthouse.co.uk/why_charthouse?SessID=dae30f99be99d0597e661f63f80582dd - 12k -

3. *As they are trustful, they believe that a great fun is possible only with a toy seen on TV.*

---

[2] All the examples that are quoted below come from the Polish part of the International Corpus of Learner English compiled at Université Catholique de Louvain (Granger 1998).

In this case we need to change the word order and replace a number of lexical items. This produces 135 hits.

Baylor University | | Public Relations | | This History Is More ...

One can **have a lot of fun** with this book: pooh-pooing ancient superstitions and cheering the champions of progress while wishing more historians had Mr. ... www.baylor.edu/pr/index.php?id=6433 - 14k -

4. *In fact, they are merely potent tools and we have it within our power to use them as we decide.*

The problem here is the verb *decide*. When we replace it with *choose* we get 22 hits.

Questions

... as "owners" of our bodies, able to "**use**" **them as we choose**. But according to the Church, our bodies belong to us in a rather different sense than this. ... www.secondspring.co.uk/christianity/sex11.htm - 14k -

5. *But for sure, our culture will undergo an intensive permeation of foreign negative values which may in fact topple Poland's fragile state of things.*

An alternative to *undergo an intensive permeation* is a passive phrase where *permeate* is a verb (582 hits).

No Escape From Philosophy

Then our lives, even in the performance of monotonous tasks, **will be permeated** by a mood arising from our conscious participation in a meaning. ... www.wvu.edu/~lawfac/jelkins/philosophy/noescape.html - 13k -

6. *The main factor behind such a situation is the fact that teenagers are not as mature as they would like to be in the eyes of the world.*

Replacing the string *factor behind such a* with *reason for this* gives us 756 hits.

Oilcrash.com: Kevin Moore: Government Unprepared For Peak Oil

Probably the **main reason for this situation** is that the Clark government has deliberately downplayed or totally ignored very pertinent facts, ... www.oilcrash.com/articles/nz_hrld1.htm - 8k -

7. *What's more, also family life is considerably suffering because of ruling television program.*

My WWW search revealed one instance of *is considerably suffering* and 114 of *is suffering considerably*. For example:

Peter Garrett : Indigenous Education Amendment Bill

... Indigenous higher education **is suffering considerably** and the numbers of students who are able to get into indigenous higher education has in some ... www.petergarrett.com.au/c.asp?id=214 - 23k -

8. *Live broadcasting on television or radio gives us fantastic possibility to witness events happening at the very moment not leaving our comfortable armchairs.*

Even if the indefinite article *a* is placed in front of the noun *possibility*, this still produces no hits of the phrase in question. Only by using *opportunity* instead of *possibility* do we get the right results (230 hits).

Updating The Wisconsin Idea

"Being asked to help with policy issues **gives us a fantastic opportunity** to work on problems that matter to everyone." Joeres has been deeply involved in ... www.cals.wisc.edu/wfsp/no-6/6-2b.html - 8k -

9. *A concrete military system would guarantee a considerable level of safety under the umbrella of its members.*

I tried *strong military alliance* and 232 hits were displayed.

Mrs. Bush's Remarks at a Troops to Teachers Event at Aviano Air ...

We boast a **strong military alliance** that protects freedom, peace and stability throughout Europe and the Trans-Atlantic region. ... www.whitehouse.gov/news/releases/2001/07/20010720-9.html - 25k -

10. *Life in modern world has become accelerated with abundant information circulating so that each day one is overwhelmed by a huge flow of words and pictures broadcast on TV, radio, or printed in newspapers and magazines.*

I did not like two things about the highlighted string. First, I felt that the definite article should have been inserted in front of the expression *modern world*. So I searched for the string *life in the modern world has* and found 43 examples.

AFS Women's Section

Everyday **life in the modern world has** become increasingly science fictional: No longer are cloning, cyberspace, nanomachines, and prostheses simply the ... www.artlore.net/ffcarchives/2002_08_01_archives.html - 14k -

Second, I wanted to simplify the verbal complex in the string. I decided to remove the verb *become* and found 315 occurrences of *life has accelerated*.

The Earth Charter: Incarnating a New Cosmology, by Paula Toner, rscj

Meanwhile, the pace of **life has accelerated**. We are exposed to more because of information technology, travel, and urban living, making it difficult for us ... www.rscj.org/news/leadership/1003earthcharter.html - 7k -

5. CONCLUSION

I hope that the above presentation has shown that the World Wide Web is a rich source of information for learners and teachers of English and that it has a distinctly noticeable practical value. A word of caution is necessary, though. Before using a phrase found on the Web one has to take a look at the site where one has found it. If the site is a native English one (they all are in the examples above) and if its register suits one's material, then one can use it as a linguistic resource with confidence.

One general (albeit tentative) conclusion that can be drawn from the survey reported on in section 3 is that corpus linguistics and data-driven instructional techniques have made no impact on the classroom methodology of foreign language teaching in the Polish context. It seems that introducing teachers and students to the idea of using the WWW as a linguistic corpus could be the first step in getting them to appreciate the benefits of methodological procedures based on general corpora.

REFERENCES

Beatty, K., 2003, *Teaching and researching computer-assisted language learning*. London: Longman.

Bley-Vroman, R., 1989, What is the logical problem of foreign language learning? In: S.M. Gass / J. Schachter (eds.), *Linguistic perspectives on second language acquisition*. Cambridge: CUP.

Dudeney, G., 2000, *The Internet and the language classroom*. Cambridge: CUP.

Granger, S. (ed.), 1998, *Learner English on Computer*. London: Longman.

Hunston, S., 2002, *Corpora in applied linguistics*. Cambridge: CUP.

Johns, T., 1994, From printout to handout: Grammar and vocabulary teaching in the context of Data-driven Learning. In: T. Odlin (ed.), *Perspectives on pedagogical grammar*. Cambridge: CUP.

Meunier, F., 2002, The pedagogical value of native and learner corpora in EFL grammar teaching. In: S. Granger/J. Hung/S. Petch-Tyson (eds.), *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam/Philadelphia: John Benjamins.

Pawley, A./F. Syder., 1983, Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J.C. Richards/R. Schmidt (eds.), *Language and Communication*. London: Longman.

Schachter, J./Tyson, A.F./Diffley F.J., 1976, Learner intuitions of grammaticality. *Language Learning* 26/1: 67-76.

Sinclair, J., 1991, *Corpus Concordance Collocation*. Oxford: OUP.

Teeler, D./ Gray P., 2000, *How to use the Internet in ELT*. Harlow: Longman.