Integrating Language Learning and Teaching with the Construction of Computer Learner Corpora

Duane Kindt
Mark Wright

INTRODUCTION

The increasing popularity of using language corpora to inform language learning and teaching (LLT) is closely related to the widespread availability of fast personal computers and increasingly sophisticated software. Though corpus linguistics is now a well-established field initiated by pioneers such as Francis (1964) and Quirk (1968), the past 40 years have seen a gradual rise in the number of published works, leading to a boom in 1998 (for example, Biber, Conrad, & Reppen, 1998; Granger, 1998c; Kennedy, 1998; Oakes, 1998; Partington, 1998; Renouf, 1998). In the wake of this boom, the interest in putting tools for corpora analysis in the hands of language teachers has also increased dramatically. What has made these tools attractive for teachers? How do they change our view of language and language learning? How can the construction of corpora be integrated with classroom activities? And how do students react to corpus-informed materials (CIMs)?

The purpose of this paper is to explore these questions. Assuming that most readers are new to working with corpora and corpus software, and to applying results to language-learning materials, we begin by briefly describing the origins of computer corpora and their applications. We offer some influences of corpus-based research (CBR) on how language is viewed before introducing an important subset of language corpora, the computer learner corpus (CLC) and elaborate on our reasons for interest in this area. This is followed by a description of our first corpus-building project along with classroom application and student reactions. To conclude, we briefly discuss the future of learner corpora construction at Nanzan University and encourage others to develop "home-grown" learner corpora of their own.

LANGUAGE CORPORA BASICS

Barnbrook (1996) defines a *corpus* (plural *corpora*) as "a sample of a larger collection of language [that] is intended to allow conclusions to be drawn about this larger body" (p. 24). If this collection of language has an index to the words in the text (Sinclair, 1991, p. 179), it is called a *concordance*. Corpora and concordances, albeit in print form, have been used in academia for hundreds of years. The first concordance of the Bible, for example, was compiled in the thirteenth century by "no fewer than 500 monks" (Tribble & Jones, 1990, p. 7). Obviously, working manually was and still is extremely time consuming; personal computers changed that. Teachers and researchers can now use specialized software to relatively quickly and accurately derive lists of lexical frequency, syntactical frequency, words in context, the degrees of relative co-occurrence of words and phrases, and a variety of other statistical information from a corpus. In turn, this information can be used to inform linguistic research, dictionary writing, and LLT in general.

There are many types of computer corpora and methods of data analysis. While Renouf (1997) notes that "it is clear that the totality of language cannot be known and so the corpus will not (unless accidentally) be a perfect microcosm of the whole" (p. 258), some corpora are large enough to be *representative* of a whole language. One example is the more than 100,000,000 word British National Corpus (BNC) (corpora and related resources are listed in Appendix 3). Others, like the Michigan Corpus of Academic Spoken English (MICASE) and the Corpus of London Teenage Language (COLT), are more specialized corpora. Besides the size distinction, corpora can differ in language, varieties of language, gender, age group, native versus non-native, mode (written or spoken), and the like.

When looking at corpus data, one important distinction is that of *raw data* and *annotated data.* According to Mallikamas (1999), raw data is plain text (words and punctuation marks without any other information) and annotated data is raw text with formatting information (page breaks, paragraphs, fonts), identifying information (author, date, genre), and linguistic information (word class, syntactic structure, discourse markers). The Brown Corpus, the Lancaster/Oslo-Bergen Corpus (LOB) and the British National Corpus (BNC) are examples of grammatically annotated corpora.

Gathering written data is relatively easy; on the other hand, "transcribing recorded speech is a very tedious and time-consuming process" (Tribble & Jones, 1990, p. 19). Though less common than written corpora, emphasis on spoken corpora has gradually increased, becoming a major focus in the mid-nineties (see Leech, Myers, & Thomas, 1995; Knowles, Wichmann, & Alderson, 1996). Examples of spoken corpora are the

Lancaster/IBM Spoken English Corpus (SEC), the Corpus of Spoken, Professional American-English, and the spoken sections of the Bank of English.

Language that attempts to mirror speech, such as TV programs, movies, plays, and the like, are not considered natural speech and fit somewhere between the spoken and written forms. This is an important distinction, especially when discussing *authenticity*, which despite criticisms (see de Beaugrand, 2000), is gradually becoming the norm in materials design (Granger, 1998a, p. 175).

A CORPUS-INFORMED VIEW OF LANGUAGE

Results of CBR have drastically changed the way many scholars view language. This is due in great part to the fact that corpora provides "the basis for more accurate and reliable descriptions of how languages are structured and used" (Kennedy, 1998, p. 88) As Zhang (2000) notes, corpora have the advantage of "providing large databases of naturally-occurring discourse so that analyses can be based on *real structures and patterns of use rather than perceptions and intuitions*" (p. 9, emphasis added). Since it is literally impossible for any one person to experience all of language, researchers and teachers are freed from relying entirely on intuition and are able to find explanations that fit the evidence, rather than adjusting the evidence to fit a pre-set explanation (Sinclair, 1991, quoted in Stevens, 1995). From this point of view, CBR is seen as providing "better descriptions of a language by embodying a view of it which is beyond any one individual's experience" (Aston & Burnard, 1998, p. 5). This wider view would improve pedagogy both by providing better reference tools and more informed choice of which lexical items and grammatical structures to include in the syllabus (Aston, 1995; Sinclair & Renouf, 1988; Willis, 1990).

There are two contradictory principles of language organization that need elaboration here—the *open choice principle* and the *principle of idiom.* Following the open choice principle, "we teach sentence frames and then students manipulate those frames in particular ways to achieve particular meanings" (Willis, 1997). The principle of idiom challenges this view:

> The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments. (Sinclair, 1991, p. 110)

The most immediately useful and accessible aspect of corpus investigation is lexical analysis, informing teachers which words and semi-preconstructed phrases students are most likely to need and thus most important to teach (Meunier, 1998, p. 28) Mindt (1996) sums up the benefit of using corpora in materials design: “Corpus-based decisions on foreign language teaching syllabuses could help considerably to bring textbooks for teaching English as a foreign language into closer correspondence with actual English” (p. 247).

As the results of corpora-based research change the way we view and teach languages, one might be tempted to see corpus-informed techniques as vying to become the next new method. This is far from the case. As Flowerdew (1993) notes, corpus-informed techniques can contribute equally well to product and process foci (p. 241). Corpus information is useful for researchers and teacher of “virtually all methodological inclinations” (Mark, 1998, p. 24). We see teachers who work with corpora as having added to their teaching repertoire. For example, a teacher who emphasizes learner-centeredness and active learning, also called discovery learning by Tribble and Jones (1990, p. 35), can exploit the learner-as-researcher nature of CIMs (Aston, 1997).

BRINGING CORPORA KNOWLEDGE TO THE LANGUAGE CLASSROOM

According to Higgins (1991) “the most valuable contribution a computer can make to language learning is in supplying…masses of authentic data….The most powerful of these tools is the concordancer” (p. 6). Using concordance software, words can be displayed in context. This display is often called the KWIC (keyword in context) function. It groups chosen words in a way that allows patterns to emerge. The output below, for example, was created by the concordance software at the Bank of English website (*Bank of English*, 2000). The keywords *what do you think about*[1] were entered to the UK spoken corpus, resulting in 91 hits, the first 12 reproduced below:

```
an important thing <ZGY> <ZGY> calcium but    what do you think about   the taking calcium supplement? <F01> I
     got to er exist haven’t they? <M01> Oh yes    what do you think about   <ZF1> the <ZF0> the increase in VAT to
  Graham. <M03> Good evening. <M0X> Erm    what do you think about   the role of TV in football er ie changing
 between er with neighbors somebody said well    what do you think about   it and it we noticed there was a
those under a cloud <M01> Mm. <M22> I mean    what do you think about   it? <M01> Well yeah I’m <ZF1> I
 <F0X> <ZGY> normally I don’t mind. <F0X>    what do you think about   that MX? <M0X> <ZF1> I <ZF0>
the big change come with MX and er Sir MX or    what do you think about   that? <F01> Well each generation of
      <F02> You would definitely have died. So    what do you think about   people like this? <M09> Well there’s
<M25> <ZF0> I think it’s terrible <M21> mate    what do you think about   that then eh? What do you think about
      mate what do you think about that then eh?    What do you think about   that? What do you think <ZG1> of
        this is the information I’m passing. Er not    what do you think about   it. Er and that takes time for them to get
<F04> Erm <M01> <ZF1> What do you <ZF0>    what do you think about   making slurping sounds when you drink
```

[1] [illegible]

Note that this is an annotated text with indicators of speakers, speaking turns, pauses, and the like. With an adequate number of hits (lines), the amount of text visible in a KWIC display is usually enough to derive meaning of the key word or phrase (Aston & Burnard, 1998, p. 7) and indicate grammatical form (Flowerdew, 1993, p. 238). These printouts can be effective for language learning, but especially if students are obtaining answers through interpretation and categorization (Gavioli, 1997, p. 87). To facilitate this, they should be introduced to some of the same processes researchers use. Thus we agree with Johns' (1991) perception that:

> "research is too serious to be left to the researchers": that the language learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data. (p. 2)

Along with KWIC data, another effective use of concordances is to obtain collocational information. Collocation is defined by Sinclair (1991) as "the occurrence of two or more words within a short space of each other in a text" (p. 170). Since "collocational knowledge is one of the things which contribute to the difference between native speakers and second language learners" (Shei, 2000), it follows that this may be an effective focus for teaching and learning. As Lewis (2000) notes:

> Learners need to notice words with words with the words with which they naturally occur. They need guidance on what can be generalized, and…the all-important gaps…*wage war* but not **wage conflict*, **wage battle*… (p. 133)

Lewis further explains that we can recognize *lexical collocations*, the combinations of open class words, that are regularly found together: *fire escape, radio station, examine thoroughly*, and the like. We also recognize *grammatical collocations*, which combine one open class word (lexical) with a closed class word (grammatical), like *aware of* and *interested in*. Collocations are often idiomatic, like *break the silence* rather than *interrupt* or *explode*, their familiarity often hiding the fact that they are idiomatic (pp. 133-5). Clearly, if students can learn from collocations related to a particular topic, their speed in language processing and production alike would approach more native-like fluency (Aston, 1995, p. 261).

Besides collocation, information from corpora software may also include *lemmatized lists*, part of speech *tagging*, *parsing*, *z-scores*, *t–scores*, and *mutual information*. Detailed explanation of these statistical methods is beyond the scope of this paper.

However, Barnbrook (1996) offers brief explanations: lemmas are related inflected forms of a word; for example in searching for *swim*, one might also want the lemmas *swam, swum, swimming* (p. 50). Tagging, as mentioned earlier, commonly adds grammatical indicators such as word class or part of speech to raw text (p. 109). Parsing, which comes after tagging, identifies words functions as parts of a clause (p. 110). Z-scores, t-scores, and mutual information are different methods of measuring relative probability of co-occurrence of words and phrases (p. 98).

Besides these and many other contributions of corpus research to the understanding of naturally occurring language, corpus research may contribute to LLT evaluation as well. Though in early stages of development, corpora might be used for test selection, test construction, and in the scoring process (Alderson, 1996, p. 253).

COMPUTER LEARNER CORPORA

Though data from native-speaker corpora is substantial, to further apply corpus-based research to LLT, Shei (2000) argues that teachers should become familiar with findings from both large native-speaker corpora (reference corpora) and target learner corpora (to show common mistakes). If comparison of learner language and native-speaker language is beneficial, it follows that emphasis should be placed on developing learner corpora. Although we agree with Tribble and Jones (1990) that “learner texts...have no place in a general corpus” (p. 18), we will show that they can and do have an important place in corpus linguistics and informing LLT.

Not surprising, the majority of studies in computer learner corpora (CLC) are with written corpora. But with the current emphasis on oral communication as the primary judge of general language ability, and even intelligence, (Nunan, 1999), efforts in gathering spoken learner data is increasing. Granger (1998a, p. 176) argues the benefits of learner corpora:

> the efficiency of EFL tools could be improved if material designers had access not only to authentic native data but also...to authentic learner data, with the NS (native speaker) data giving information on what is typical in English, and the NNS (non-native speaker) data highlighting what is difficult for learners in general and for specific groups of learners. (p. 176)

Probably the best known of learner corpora is the International Corpus of Learner English (ICLE). The purpose of this corpus is to allow for the systematic comparison of English learners of 14 nationalities. It consists of academic essays, mostly from students in their third year, that are approximately 500 words long and lend themselves to analyses of coherence and cohesion (Granger, 1998b, p. 10).

While a corpus of various L1 backgrounds would help in the discovery of difficulties particular to certain groups or to all groups, a homogenous learner corpus will “reveal features of non-nativeness” (Meunier, 1998, p. 19). Several teachers have initiated projects related to the collection of Japanese learner corpora. Mark (1998) has built “a 300,000 word corpus of Japanese university student English writing” which has been “contrasted with...a 60,000 word corpus of British spoken English” (p. 24). Tono (1999) is working on the Japanese EFL learner corpora (JEFLL). The purpose of this project is to see how different modes of teacher feedback would affect EFL learners’ performance in essay writing.

Though many aspects of corpus design mentioned earlier apply to learner corpora, we will focus on two in particular. In learner corpora errors need to be indicated (tagged). The ICLE indicates errors and corrections by first tagging the error and then putting the correction between dollar signs after the error, as indicated by the excerpt below:

```
Actually $In fact$ studies carried out by eminent linguists have
proved that the ideal (FS) adge $age$ for learning a foreign
language is between 3-10 years. (CLC) Of course $0$ no consensus
(GVAUX) could be reached $has been reached$ about this subject as
```

(Granger, 1998a, p. 183)

The zero between dollar signs indicates that what comes before it should be omitted. Deciding tagging procedures in a major issue in CLC.

Another important design consideration is size. The appropriate size depends upon its use and cannot be predicted. This means that although the purpose of a corpus may be clear, as data is gathered insights may emerge that change “The whole point of assembling a corpus is to gather data in quantity,” though “the size…tends to reflect the ease of difficulty of acquiring the material (Sinclair, 1995, p. 21, quoted in Granger, 1998b).

Though transcription system and size are important considerations, for the relatively work-intensive corpus collection to be successful—as with any innovation—there needs to be a clearly-defined, beneficial purpose. In promoting purposeful use of corpora in the classroom, Franca (1999) notes that working with corpora of spoken learner data helps students to:

- strike a balance between oral fluency and accuracy and to become critically aware of their own production.
- become aware of features present in conversational interaction during problem-solving activities.
- acquire greater confidence in more spontaneous oral interactions whilst at the same time maintain control over the features which help sustain conversational interaction.
- have realistic targets in terms of their own interlanguage production, comparing their production with that of more proficient non-native L2 speakers. (p. 116)

Thus, it seems clear that concordance-based comparisons frequently highlight erroneous patterns in learner production (Granger, 1998a, p. 178). If it is true that highlighting error patterns can increase students' awareness of the difference between their own problematic speech patterns and correct patterns, moving them closer to native-like speech, then this would be an appropriate purpose indeed, and one we decided to pursue.

COLLECTING AND USING CORPORA DATA

The learner corpus at Nanzan University is being developed to inform current syllabus design and classroom pedagogy. This decision was supported by the "growing use of smaller, homemade corpora…to create classroom materials and exercises" (Aston, 1997). The database is intended for use with popular word-processing programs (e.g.: Microsoft Word ("Word98," 1998); Nisus Writer ("Nisus Writer," 1994)) and programs particular to analyzing corpora (e.g.: WordSmith (Scott, 1996, 98); CONC (Thomson, 1992)). These word-processing programs can assist in the detection of spelling and grammatical errors, and also in finding derivational errors, inflectional errors, word coinages and blends (Meunier, 1998, p. 26).

Collecting data for a written learner corpus is relatively easy. Data can be downloaded from the corpora posted on the Internet (Tono, 1999) or pasted from students' work already in electronic form (Mark, 1998). In gathering a corpus of students' oral production there is added difficulty in recording and transcription, and this means ensuring that data collection is done ethically (students gave permission for their transcripts to be used in class).

Making sure that research does not interfere negatively with the normal conduct of a class is important as well. Like Franca (1999), we see the construction of a learner corpus "as a gradual process based on the learners' normal classroom activities (p. 116). Thus, for the initial stages of data gathering, nothing extraordinary was asked of students. Even the procedure for submitting the transcription of their conversations in electronic form was familiar to students. Supported by Bauman's (1998) view that as communication teachers we cannot ignore the “increasing importance of electronic communication,” we asked students to submit assignments via email.

The effect of task-based instruction on students’ learning is still a hotly debated topic (Lewis, 1993; Long, 2001; Skehan, 1996). In considering the task for data collection, we agreed with Bruton (1999) that tasks are often “very artificial” and sometimes overly influenced by the teacher’s desire to get students using particular lexical units or grammar forms, sometimes brought about by an ironic overemphasis on corpus evidence (p. 26). In fact, we admit that there is a concern whether corpora should inform task, or task should inform corpora. For us, the two should inform one another, especially at the beginning stages of corpus building.

With this in mind, we invited students to work together to decide the recording topic. Though we knew the topic, we could not predict what insights the data might give us. We followed Scholfield (1995) in taking a “hypothesis-finding” approach, gathering data from a particular “language activity in the classroom…just to see what emerges” (p. 24). We anticipated that while building a learner corpus we would be able to compare the students’ spoken English to a native-speaker corpus and discover some aspects of language learning beneficial to students.

The corpora-informed activity

The students in the study are English majors at Nanzan University. They meet Monday, Wednesday, and Friday for 45 minutes. There are two classes of approximately 20 students; three students in each class are male. They are familiar with deciding conversation topics as a class and preparing for either audio- or videotape conversations. To prepare they often study example conversations from former students and make conversations cards (Kindt, 2000). After gathering transcriptions from the first conversation, one class received the CIM, *class with* (CW), and one that did not, *control class* (CC). To do this we followed the procedure listed below:

1) At home, **students prepare** to ask *What do you think…* questions.
2) When returning to class, students practice the conversation once before **recording a 5-minute conversation**.
3) For homework, **students transcribe the conversations** without any corrections except spell checking, **and email the transcriptions to the teacher**.
4) As the teacher receives students’ transcriptions, he converts them to a text file in Word98, **highlighting spelling errors (underlined) and grammar errors (uppercase)** and **printing the files.**
5) After the assignment deadline, **the teacher pastes all submissions into one file**.

6) **Concordance software** (Conc and WordSmith) **are used to analyze students' language,** the results informing learning materials that focus on those errors.
7) In the next class meeting, **the CIM is given to one class (CW)**.
8) **The original transcription is given** to both the class with the CIM and the class without (CC).
9) In the next class, **students make another *What do you think…* recording.**
10) **Repeat** numbers 2 through 8.

One of the most prohibitive aspects of gathering spoken data is “the time necessary for transcription and adopting and applying a transcription convention” (Barnbrook, 1996, p. 33). But since the purpose of our study was to discover some errors which are common for a particular group of students when asking and answering *What do you think…* questions, a corpus that easily shows error patterns is ideal. Thus, we decided to underline spelling errors, highlight grammar errors, and print out the transcriptions before putting the transcriptions into the corpus. All grammar errors were indicated by putting any incorrect or unnecessary words in uppercase letters (see Appendix 1). When the error was difficult to indicate, as in an omission, the words on either side of where the word should be are in capital letters.

This pattern of highlighting learner error is similar to the form used in the students’ textbook Kindt (2000). It is a procedure that makes it “easy to distinguish from the original words” (Barnbrook, 1996, p. 110) and provides students with negative linguistic evidence (Rohde & Plaut, 1999). The benefit is that students can more readily find commonly made mistakes and focus their energies on avoiding those mistakes. What is more, transcripts of this type can be quickly transformed into classroom material; the original versions could be used for students to correct as language-learning activities. Initial reactions to finding errors within students’ *own* transcriptions, however, lead us to believe that this process is more effective.

After receiving the first email messages, we pasted the transcribed conversations into a Word98 file. This allowed us to note errors using the spelling and grammar check functions. In Word98, the spell-checker function underlines errors in red while grammar errors are underlined in green. Since these colored underlines cannot be printed, nor would they appear in corpus software, we changed the grammar errors to uppercase letters, and spelling errors were underlined manually.

Care was taken to avoid errors, but only a reasonable degree. Jones (1997) notes, “even after numerous readings there will still be errors and words that were not comprehensible. That is the nature of working with spoken language" (p. 149). The degree of error when conducting linguistic research must be minimized, but in looking for basic patterns to helps students find and improve a few errors, a perfect transcription was not the emphasis.

Though all students were given their transcriptions for correction, only one class was given a worksheet derived from corpus information. This was to create pseudo-experimental situation that would hopefully provide comparable results and useful insights. Initially, we planned to use corpus software to discover common error patterns in students' transcriptions. Learner corpora when matched with native-speaker corpora are particularly useful for Contrastive Interlanguage Analysis (CIA) (Granger, 1998b, p. 12). Initially, we thought the correct examples would be provided by native-speaker corpora, but because of our support of near-peer role models (Assinder, 1991; Murphey, 1998)—a concept which students where introduced to in their textbook (Kindt, 2000)—and because many of the common mistakes appeared in correct form elsewhere *in the same learner corpus*, we decided to use students' sentences exclusively. At the same time, we admit this is inappropriate for teaching some structures or for comparing native-like frequencies and collocation.

The class that received the CIM was asked to try to formulate a rule of their own from the concordance lines with errors when compared to those without and try to explain their rule to their partner (see Appendix 2). Toward the end of class they were asked to correct their conversation transcripts at home. The class that did not receive the CIM corrected their transcriptions in class. Both groups were asked to prepare again to record conversations on the same theme but with different partners in the next class.

After recording and transcribing the second *What do you think…* conversation, students again submitted their assignment via email. These were printed out without any highlights and returned to students the following class period. This was to explore the benefit of transcriptions with and without highlights. After correcting their conversations in class, both classes were given a questionnaire related to the two recordings.

PRELIMINARY RESULTS AND STUDENT REACTION

Table 1 shows information related to word counts from the *What do you think…* conversations.

	Think 1 (total)	Think 1 (avg)	Think 2 (total)	Think 2 (avg)	Change (avg)
CC	8,480	424	8,275	414	-10
CW	5,967	332	6,836	380	+48

Table 1: Statistics from the *What do you think…* conversations

At first glance, one might conclude that CW obviously benefited from the CIM, increasing their average word counts by 48, as opposed to CC's reduction by an average of 10 words per conversation. Perhaps because CW had focused on how questions were formed for this topic, they may have been better able to "get the question right" and then switch to a focus on fluency in their answers. Word production alone, however, is not a sufficient indicator of progress and relying solely on word counts is problematic. In fact, Doughty and Varela (1998) have cautioned that a greater focus on form can draw students' attention to accuracy, reducing their output. And we can only guess what made CW's average lower in the first place, especially when they have consistently outperformed CC in previous recordings. In fact, this was the reason for giving the CIM to CW in the first place—because they seemed better able to deal with the greater work load.

After students discussed their second transcription, they were given a questionnaire concerning the project. CW was asked to indicate the following:

How much did the following influence your learning? ***very little*** ***a lot!***

1. Trying the new email system for the first time with a 2-minute conversation.................1—2—3—4—5—6
2. Practicing for the **FIRST** “What do you think…” conversation.....................................1—2—3—4—5—6
3. Recording the **FIRST** “What do you think…” conversation ....1—2—3—4—5—6
4. Writing the **FIRST** “What do you think…” email transcription.....................................1—2—3—4—5—6
5. Correcting errors in the print with correct (O) and incorrect (X) lines ....1—2—3—4—5—6
6. Correcting errors in your **FIRST** transcription ....1—2—3—4—5—6
7. Practicing for the **SECOND** “What do you think…” conversation.................................1—2—3—4—5—6
8. Recording the **SECOND** “What do you think…” conversation......................................1—2—3—4—5—6
9. Writing the **SECOND** “What do you think?” email transcription..................................1—2—3—4—5—6
10. Correcting the errors in your **SECOND** transcription ....1—2—3—4—5—6
11. Comparing the first and second conversations...1—2—3—4—5—6

Note that item #5 was not included on the questionnaire for CC. Table 2 indicates the results:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
CW	4.47	5.18	5.25	4.56	5.00	5.47	4.76	4.94	4.59	5.18	4.82
CC	4.65	5.59	5.63	5.21	—	5.32	5.63	5.74	5.26	5.21	5.16

Table 1: Statistics from the *What do you think…* project questionnaire

CC rated every activity higher, except #6 (correcting errors in the first transcription), which was also CW's highest rating. It would seem that the corpus-informed handout is contributing positively to how students rate later work with individual transcriptions. Students also responded in writing to these items:

1. Would you prefer submitting your transcriptions by email, on a form, as a summary, or some combination of these?
2. Explain which is better for you: the first print out, with underlined spelling errors and uppercased grammar errors, or the second print out, simply printed from your email message?

Though students' comments give a wealth of data too large to reproduce here, including a few excerpts may best represent their overall reaction.

Selected student comments concerning the submission system

Today, we live in computer society. So, this is suitable. It's new for us. In conclusion, I like this system.

I like e-mail. I have my own computer but there are some students who don't have it and I can't say which is better.

I prefer writing transcription to typing and sending e-mail because I don't have a personal computer.

I like this e-mail system. I can check my conversation twice at all (transcribing, typing).

I think I should do the transcription as early as I can after the recording. It doesn't take much time if I do so.

It took a lot of time. My opinion is 5 minute conversation is too long for transcribing and typing. My partner couldn't really understand what I said and couldn't transcribe correctly…so we really can't be sure about our own mistakes.

I completed the second one more easily because I got used to the system, I think. True, the more we repeat the system, the more we'll get used to it.

It took me about half an hour to do transcription and typing. For me, the first transcription was better, but I think it's important to find my mistakes by myself, too. Well, but highlighted transcription tells me mistakes that I wouldn't recognize.

Selected student comments concerning the printout type

I prefer transcription with uppercase letters to show grammar errors. That's because I can't find all the mistake as I don't notice the mistake. I mean, it's more important to correct mistakes than to find the mistakes, in my opinion.

I really like the transcription with uppercase letters…[because] we can't know the mistakes that Japanese often make just among Japanese students.

In my opinion, we can get much more things from transcription without hints.

I'd like you to show grammar errors and underline spelling errors by changing uppercase letters. This is because I can notice my mistakes and learn it. If I don't know the parts of mistakes, I might continue use them in wrong way.

It is so difficult to find grammar mistakes. I'm always making grammar mistakes unconsciously and, to be honest, I've not understood how to use "the" completely. I think

> *the purpose of transcription is to correct our broken English. If I can’t correct mistakes precisely, it is useless to do transcription.*

There was overwhelming support for returning highlighted transcriptions, though preferences of submission system differed. Several of those who preferred the e-mail system thought the work was too hard, often due to less familiarity with computers or difficulties in access.

CW also answered this question about the corpus-informed handout:

> What do you think of learning English patterns with lines from you and your classmates’ transcriptions?

The following are some representative reactions:

> **Selected student comments concerning the corpus-informed handout**
>
> *It was a good material for studying “about.” Besides, examples were from the classmates’ transcription, so they were natural because they had actually been said. I liked the handout because it showed both correct and incorrect patterns.*
>
> *The print with correct and incorrect lines helped me so much. I could understand how to use “a” (singular form) and “~s” (the plural form).*
>
> *The print was helpful for me to correct my mistakes. I could talk more correctly in the second conversation.*
>
> *I haven’t thought about what comes after What do you think about~ phrase. I wasn’t using it properly.*
>
> *It was help me to understand my tendency of making mistakes and make a greater effort to ask my “What do you think” questions in the next conversation. I think that review is important to make our English better and to learn from our mistakes.*
>
> *It was difficult to understand the purpose of the print…until [the teacher] explained it in detail. If such prints are handed out more, I can know the points I should be careful.*

Students in CW unanimously supported the use of the CIM. Besides reasons mentioned above, this may also be due to the novelty of the activity. The comment about “such prints” being handed out more definitely encourages further efforts to use corpora in creating materials.

DISCUSSION

Though more time consuming than getting word counts, CONC software’s KWIC function can be used to compare the number of errors of a certain structure between students or classes. In the first conversation, for example, CC had 25 correct and 8 (24%) incorrect instances of *what do you think about* compared with 16 and 8 (33%) in CW. In the second conversation the results were 26 and 5 (16%), and 20 and 4 (17%), respectively. These figures were obtained by listing *what do you think about* for each class and counting the lines with errors. This gives the indication that both classes improved, but CW had a greater improvement, 16% less errors than CC’s 8% improvement. These results, of course, cannot be taken as hard evidence. Besides comparing only class averages and not individual achievement (which is also possible), students were influenced by practice, may have used more *what do you think of* questions, or simply avoided questions that they had not prepared well for. Whatever the case, the results are promising and the 30 minutes used to create this lists gave us a rich list of questions students are asking, a valuable resource for subsequent lessons, clearly a benefit of integrating LLT with the construction of a learner corpus.

Interesting results were also obtained by running the corpora through WordSmith. Using the wordlist function, we could find the frequency of words of a particular length and average word length of the conversations. The average word lengths for all the conversations were statistically the same: 3.68 and 3.65 for CC’s first and second conversations and 3.61 and 3.63 for CW. When looking at the length of 10-letter words, however, CC had 94 in the first conversation and 101 in the second. The number for CW was 31 and 87. As in the case of error reduction, this information can be misleading, but it would seem that something between the first and second recordings helped CW make noticeable improvement in many respects.

Though obviously in the very beginning stages of corpora development, initial results of integrating LLT with learner corpus construction are promising. Student response was generally positive, as was our first attempt at using learner corpora-informed materials in the language classroom. We came up with some creative techniques for using learner corpora (we have yet to read studies of errors in KWIC materials being supported by correct forms within the same learner corpus), and were able to gather material for a spoken learner corpus at the same time.

There is no question that formatting and annotating the mini-corpus in our study took the greatest amount of time and effort. We agree with Meunier (1998) that the “automatic flagging of errors would greatly facilitate the error annotation of learner corpora" (p. 27). As the future brings increasingly sophisticated technology, the work required to annotate learner corpora and create appropriate materials will be reduced.

Steps have been taken to begin the collection of student writing samples. As the direction of the corpus becomes clearer, we hope to make the raw data available to teachers, possible in the form of a CD-ROM. This would allow teachers to use the data to support their own classes and interests. Though teacher intuition is a valuable, and necessary, aspect of LLT, using a learner corpus to support the improvement of pedagogy and materials seems both timely and practical.

Though many questions related to the effect of CIMs on students' learning have yet to be answered, having a relatively large sample of students' spoken data in simple electronic form is a tremendous benefit. Though we do not argue that computers should replace teachers, the humanistic aspect of LLT being of central importance (Underhill, 1989), we do believe that teachers can be more effective when confident that the language being introduced is grounded in empirical evidence.

Though corpora-informed LLT is no panacea, as a way "help to develop new pedagogical tools and classroom practices which target more accurately the needs of the learner" (Granger, 1998b, p. 12), CLC is very promising. As Hadley (1999) notes in his study of data-driven learning with beginning students at a Japanese university, "[t]he learning opportunities provided by this approach should not be ignored." With the wealth of information already available, it is somewhat surprising that more teachers are not bringing corpora-informed materials to their classrooms. Tribble (2000) offers two possible reasons for this: "[T]eachers work in an increasingly pressured environment[, and]…there is still a high level of techno-fear" (p. 32). Understanding that corpus-based materials support efforts in increasing inductive learning, hypothesis testing, and learner autonomy (Aston, 1995; Ketteman, 1995), teachers may take the initiative to experiment with building corpora of their own and using materials based on them in their classrooms.

Acknowledgments

The authors would like to thank other members Audio-Visual-Corpora Research (AVCR) group: Prisca Molotsi and Gabriel Yardley for their support, Robert Croker for his assistance with corpora design, and Tim Murphey for the many years of directing the Audio-Visual-Corpora Research (AVCR) group.

This research was supported by generous Pache-IA grants from Nanzan University.

References

Alderson, J. C. (1996). Do corpora have a role in language assessment? In J. Thomas & M. Short (Eds.*), Using Corpora for Language Research* (pp. 248-59). London: Longman.

Assinder, W. (1991). Peer teaching, peer learning: One model. *ELT Journal*(*45*), 218-229.

Aston, G. (1995). Corpora in language pedagogy: Matching theory and practice. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics* (pp. 257-270). Oxford: Oxford University Press.

Aston, G. (1997). Involving learners in developing learning methods: Exploiting text corpora in self-access. In P. Benson & P. Voller (Eds.), *Autonomy and independence in language learning* (pp. 204-214). London: Longman.

Aston, G., & Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Bank of English. (2000). Available: http://titania.cobuild.collins.co.uk/index.html [2000, 02.14.01].

Barnbrook, G. (1996). *Language and Computers: A practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.

Bauman, J. (1998). *Using E-mail with your students*. The Language Teacher Online. Available: http://langue.hyper.chubu.ac.jp/jalt/pub/tlt/98/feb/bauman.html [2000, November 11,].

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.

Bruton, A. (1999, March 29). *Task-based instruction and its closest relations.* Paper presented at the 33rd International IATEFL (International Association of Teachers of English as a Foreign Language) Annual Conference, Herriot-Watt University, Edinburgh, Scotland.

de Beaugrand, R. (2000). *Large Corpora and Applied Linguistics: H.G. Widdowson versus J.McH. Sinclair*. Available: http://beaugrande.bizland.com/WiddowSincS.htm [2001, February 16].

Doughty, C., & Varela, E. (1998). Communicative focus on form. In C. Doughty & J. Williams (Eds.), *Focus on Form in Second Language Acquisition* (pp. 114-138). Cambridge: Cambridge University Press.

Flowerdew, J. (1993). Concordancing as a tool in course design. *System, 21*, 321-344.

Franca, V. B. (1999). Using student-produced corpora in the L2 classroom. In P. Grundy (Ed.), *IATEFL 1999 Edinburgh Conference Selections* (pp. 116-117). Whitstable: IATEFL.

Francis, W. N. (1964). *A standard sample of present-day English for use with digital computers*. Providence,: Brown University.

Gavioli, L. (1997). Exploring texts through the concordancer: Guiding the learner. In A. Wichmann & S. Fligelstone & T. McEnery & G. Knowles (Eds.), *Teaching and Language Corpora* (pp. 83-99). London: Longman.

Granger, S. (1998a). The computer learner corpus: A testbed for electronic EFL tools. In J. Nerbonne (Ed.), *Linguistic Databases* (pp. 175-188). Stanford, CA: CSLI Publications.

Granger, S. (1998b). The computer learner corpus: a versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on Computer* (pp. 3-18). London: Longman

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

Hadley, G. (1999). *Sensing the winds of change: An introduction to data-driven learning*. Available: http://web.bham.ac.uk/johnstf/winds.htm [2000, November, 19].

Higgins, J. (1991). Fuel for learning: The neglected element of textbooks in CALL. *CAELL Journal*, *2*(2), 3-7.

Johns, T. (1991). Should you be persuaded—two samples of data-driven learning materials. In T. Johns & P. King (Eds.), *Classroom Concordancing* (Vol. 4). Birmingham: Center for English Language Studies.

Jones, R. L. (1997). Creating and using a corpus of spoken German. In A. Wichmann & S. Fligelstone & T. McEnery & G. Knowles (Eds.), *Teaching and Language Corpora* (pp. 145-156). London: Longman.

Kennedy, G. D. (1998). *An Introduction to Corpus Linguistics*. London; New York: Longman.

Ketteman, B. (1995). *On the use of concordancing in ELT*. Available: http://gewi.kfunigraz.ac.at/~ketteman/conco.html [2000, November 20].

Kindt, D. (2000). *Don't forget your SOCCs!* Nagoya: Sankeisha.

Knowles, G., Wichmann, A., & Alderson, P. (1996). *Working with Speech: Perspectives on research into the Lancaster/IBM spoken English corpus*. London: Longman.

Leech, G., Myers, G., & Thomas, J. (Eds.). (1995). *Spoken English on Computer: Transcription, mark-up and application*. New York: Longman.

Lewis, M. (1993). *The Lexical Approach: The state of ELT and a way forward*. Hove: Language Teaching Publications.

Lewis, M. (Ed.). (2000). *Teaching collocation: Further developments in the lexical approach*. Hove, England: Language Teaching Publications.

Long, M. (2001). *Task Based Language Teaching*. Oxford: Blackwell.

Mallikamas, P. (1999). *Applications of corpora in language teaching*. Available: http://www.thaitesol.org/bulletin/1201/120101.html [2001, February 19].

Mark, K. (1998). *A Japanese learner corpus and its uses.* Paper presented at the IATEFL 1998: Manchester conference selections.

Meunier, F. (1998). Computer tools for the analysis of learner corpora. In S. Granger (Ed.), *Learner English on Computer* (pp. 19-37). London: Longman.

Mindt, D. (1996). English corpus linguistics and the foreign language teaching syllabus. In J. Thomas & M. Short (Eds.), *Using Corpora for Language Research* (pp. 232-247). London: Longman.

Murphey, T. (1998). *Motivating with Near Peer Role Models*. Available: www.ic.nanzan-u.ac.jp/~mits/pages/nprm.html [2000, December 27].

Nisus Writer. (1994). Solana Beach, CA: Nisus Software, Inc.

Nunan, D. (1999). *Second Language Teaching and Learning*. Boston: Heinle & Heinle.

Oakes, M. P. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Partington, A. (1998). *Patterns and meanings: using corpora for English language research and teaching*. Amsterdam ; Philadelphia: J. Benjamins Pub.

Quirk, R. (1968). *Essays on the English language, medieval and modern*. Bloomington: Indiana University Press.

Renouf, A. (1998). *Explorations in corpus linguistics*. Amsterdam: Rodopi.

Renouf, A. (1997). Teaching corpus linguistics to teachers of English. In A. Wichmann & S. Fligelstone & T. McEnery & G. Knowles (Eds.), *Teaching and Language Corpora* (pp. 255-266). London: Longman.

Rohde, D. L. T., & Plaut, D. C. (1999). Language acquisition in the absence of negative evidence: how important is starting small? *Cognition, 72*, 67-109.

Scholfield, P. (1995). *Quantifying Language*. Clevedon: Multilingual Matters.

Scott, M. (1996, 98). WordSmith Tools.

Shei, C. (2000). *Collocation, Learner Corpus, Language Teaching*. Available: http://www.dai.ed.ac.uk/homes/shei/collocation/top.html [2000, November 21].

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J., & Renouf, A. (1988). A lexical syllabus for language learning. In R. Carter & M. McCarthy (Eds.), *Vocabulary and Language Teaching* (pp. 140-160). London: Longman.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Lingusitics, 17*, 38-62.

Stevens, V. (1995). Concordancing with language learners: Why? When? What? *CAELL Journal, 6*(2), 2-10.

Thomson, J. (1992). Conc (Version 1.70 beta): The Summer Institute of Linguistics.

Tono, Y. (1999). *JEFLL Corpus: Description*. Lancaster University. Available: http://www.lancs.ac.uk/postgrad/tono/ [2000, January 17].

Tribble, C. (2000). Practical Uses for language corpora in ELT. In P. Brett & G. Motteram (Eds.), *A Special Interest in Computers: Learning and teaching with information and communications technology (ICT)* (pp. 31-42). Eynsham, UK: Information Press.

Tribble, C., & Jones, G. (1990). *Concordances in the classroom*. Harlow: Longman.

Underhill, A. (1989). Process in Humanistic Education. *English Language Teaching Journal, 43*, 250-256.

Willis, J. D. (1990). *The Lexical Syllabus*. London: Collins.

Willis, J. D. (1997). *Lexical phrases in syllabus and materials design*. The Language Teacher Online. Available: http://langue.hyper.chubu.ac.jp/jalt/pub/tlt/97/sep/willis.html [2000, November 20].

Word98. (1998). Microsoft Corporation.

Zhang, W. (2000). Corpus Studies: Their Implications for ELT. *IATEFL Issues* (152), 9–10.

Appendix 1

Example transcription

T: Think?
N: Takako Nishimura [pseudonym; female]
C: OC4b
D: 11/24/2000

Takako: Good morning.
Yasuko [pseudonym; female]: Good morning.
T: What do you think about Internet shopping?
Y: Ah, Internet shopping. To tell the truth, I have never bought anything.
T: Through… by USING INTERNET?
Y: Yeah.
T: Me, too. But why DIDN'T you use Internet shopping?
Y: I think it's DOUBTFUL. Have you?
T: I haven't bought anything ON INTERNET because I worried about…I'm worrying that that might cause some problems.
Y: Yeah, we have many problems.
T: For example, EVEN if we pay FOR the bill, but we can't get the goods.
Y: I'M THINK OF IT.
T: I think Internet is so useful when I try to find some information for research papers or reports, something like that. But sometimes I can't trust the Internet.
Y: Me, too. I think THE famous COMPANY IS TRUSTED.
T: Really?
Y: Yeah, for example, Yoji-ya.
T : Yoji-ya? I don't know that.
Y: You know, the paper which…
T: Yeah, I remember that.
Y : Or Victorias Secrete. IT a maker FOR underwear. Their underwear is very, very cute and cheap. But we don't have the shop in Japan. In Yoji-ya, we can get…
T: Kyoto?
Y: Yeah, IN ONLY Kyoto. INTERNET is useful for SUCH SHOP.
T: But when we buy something on the Internet, we must put our credit card number or bank number, password on the Internet. So I'm afraid.
Y: Maybe we can CHOOSE WHICH… when we get the goods, then.
T: After we get the goods?
Y: So you can use the postman. Not post man…
T: Maybe you mean the WORKER FOR A POST OFFICE.
Y: Yeah, post office or…
T: Delivery service?
Y: Yamato Takkubin or something like that. Maybe you can CHOOSE. I have never done it. I don't know but maybe you can CHOOSE.
T: I wanna talking about INTERNET more, more. I also... I often use the cite of fortune telling.
Y: Oh really.
T: Have you USE the fortune telling cite?
Y: Not. No.
T: No? Don't you like fortune telling?
Y: I don't, maybe. I use it for e-mail.
T: Oh really?
Y: I always use it…
T: It's time. So thank you very much.
Y: Thank you.

Appendix 2

Corpus-informed material (instructions and 3 of 6 items)

Common errors from “What do you think…

Look at the lines with errors (X) and the lines without (O). Can you make your own rule? Tell a classmate! ☺

1.

X	Ah...I THINK WHAT sort of, no I am not thinking	**ABOUT**, because NEXT I FOURTH GRADE. And What
X	many PROBLEM to think about..T: TALKING	**about** and decide and make a law. Y: So, I think their
O	studying. So... it's worrying...I'm worrying	**about it** now. What do you think of it? S : I want to
O	paper. But it'll be HOLD... I'm really sorry	**about it.** B: Oh, really. A: But if it is HOLD, I think it
O	do you think? Let me know your opinions	**about it.** M: Well, I've already told you...I have already

2.

X	P: Ok. My topic is... No.1, "What do you think	**ABOUT JAPANESE** Prime Minister, Mr. Mori? M: Oh,
X	of Christmas day. Y: Thank you. Can I talk	**about NEXT topic**? T: OK. Y: HOW do you think about
X	but maybe you can choose. T: I wanna TALKING	**about INTERNET** more, more. I also... I often use the
X	Yeah. M ; Yeah. R ; Next topic. What do you think	**ABOUT COLLECTION** of garbage according to type?
X	B: Yeah. But I also...we also have to think	**ABOUT ENVIRONMENT**. A: Yeah! B: Because MANY
O	M: Yeah. The next question's...what do you think	**about the exchange students**? E: Exchange students
O	E: OK. M: FIRST one is what do you think	**about the attitude** of Japanese students in classes
O	A: Hello. B,C: Hello. A: What do you think	**about the Internet**? C: It's useful. A: Do you often use it
O	E: Yeah. M: NEXT question is, What do you think	**about this class**? E: This class? M: Yeah. E: I like it. I
O	S: Good morning. M: Can I ask your opinion	**about my topic**? Ah, first, what do you think about

3.

X	M: ...so I can't. OK, next. What do you think	**about New Year CARD**? S: New Year CARD? M: New
X	a bit. Well, do you like sports? You asked me	**about sports PLAYER** a lot. P: Yes, I always watch
X	with family at first. I BECAME TO LIVE alone	**about one and half YEAR** ago. I was a lonely. But I'm
X	political policy? Y: So, I want them to talk	**about other much more important THING**. T: That's
X	Aiko: Tomoko: takahito A: What do you think	**about BIRTHDAY**? T: WHEN YOUR birthday? A
X	T: Yes. While I'm chatting, I don't... don't care	**about the... around, STUDENT** AROUND. But when I'm
O	a question. Aiko: Yes. Tomo: What do you think	**about winter sports**? Aiko: Winter sports? I don't
O	about my topic? Ah, first, what do you think	**about people** who are sitting on a floor or ground
O	you TO live alone. Um...I had no chance to think	**about my parents**. M ; Ah. R ; And now so I often think
O	is not BALANCE. H: I see. J: What do you think	**about many girls** and LITTLE boys? H: I think this
O	AGAINST the... A: Yeah.. B: We have to think	**about those people**. A: Yeah. I READ THE PEOPLE...ah

APPENDIX 3: Corpora-related resources

Although in no way exhaustive, below are several resources for those interested in learning more about computer language corpora and its uses in language research and language teaching. Access to websites listed are available through the **Nanzan AVCRG website** at: http://www.ic.nanzan-u.ac.jp/~dukindt/pages/NCLC.html

MIXED CORPORA

American National Corpus (ANC) available at: http://www.lsa.umich.edu/eli/micase/micase.htm
British National Corpus (BNC) available at: http://info.ox.ac.uk/bnc/
COBUILD-DIRECT and the Bank of English available at: http//titania.cobuild.collins.co.uk
The International Corpus of English (ICE) available at: http://www.mpi.nl/world/ISLE/overview/Overview_ICE.html

WRITTEN CORPORA

The Brown Corpus, The Lancaster/Oslo-Bergen Corpus (LOB), and **Australian Corpus of English (ACE)**, among others, are available from **ICAME** at: http://www.hd.uib.no/corpora.html
Corpus of English-Canadian Writing at Queen's University.
Corpus of Written British Creole available at: http://www.ling.lancs.ac.uk/staff/mark/cwbc/cwbcman.htm

SPOKEN CORPORA

Corpus of Spoken, Professional American-English available at: http://www.athel.com/cspa.html
CHILDES available at: http://www.athel.com/cspa.html
Longman-Lancaster Corpus (LLC), the **Bergen Corpus of London Teenage Language (COLT),** and the **Lancaster/IBM Spoken English Corpus (SEC)**, among others, are available from **ICAME** at http://www.hd.uib.no/corpora.html
Michigan Corpus of Academic Spoken English (MICASE) available at: http://www.lsa.umich.edu/eli/micase/micase.htm

LEARNER CORPORA

Cambridge Learners Corpus (commercial site) available at: http://uk.cambridge.org/elt/reference/clc.htm
International Corpus of Learner English (ICLE) available at: http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/cecl.html
Japanese EFL Learner Corpora (JEFLL) and Yukio Tono's corpus linguistics and SLA page available at: http://www.lancs.ac.uk/postgrad/tono/
Kojiro Asao's Corpus of English by Japanese Learners available at: http://www.lb.u-tokai.ac.jp/lcorpus/
Longman Learners' Corpus (commercial site) available at: http://www.longman-elt.com/dictionaries/corpus/lclearn.html
Samantha Spelling Error Corpus available at: http://kgaikoku.kj.yamagata-u.ac.jp/KOKUSAI/Samantha-top.html

SOFTWARE

CONC (for Macintosh computers) available at: http://www.sil.org.computing/conc
MonoConc Pro available at: http://www.athel.com/mono.html
WordSmith Tools available at: http://www.liv.ac.uk/~ms2928/homepage.html
WCONCORD available at: http://www.ifs.th-darmstadt.de/sprachlit/wconcord.htm
Wordcruncher available at: http://www.wordcruncher.com/

British National Corpus Sampler (CD-ROM) from Oxford University Computing Services.
Collins-COBUILD on CD-ROM from HarperCollins.
Collins-COBUILD English Collocations on CD-ROM available from HarperCollins.

TEXT SOURCES

Canadian Broadcasting (with sound files) available at: http://cbc.ca/onair/
Oxford Text Archive available at: http://ota.ox.ac.uk/
Project Gutenburg available at: http://www.promonet/pg/

University of Virginia Electronic Text Center available at: http://etext.lib.virginia.edu/

WEBSITES

Cathy Ball's Corpora and Concordancing Tutorial available at: http://www.georgetown.edu/cball/corpora/tutorial.htm

Mike Barlow's corpus linguistics website: http://www.ruf.rice.edu/~barlow/corpus.html

The Centre for English Corpus Linguistics (CECL) available at: http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/cecl.html

Collins-COBUILD available at:

Chris Greaves' Web Concordancer available at: http://vlc.polyu.edu.hk/scripts/concordance/WWWConcapp.htm

Tim John's Data-Driven Learning Library available at http://sun1.bham.ac.uk/johnstf/ddl_lib.htm

Tony McEnery and Andrew Wilson corpus linguistics: http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm