AN EVALUATION OF INTERMEDIATE STUDENTS' APPROACHES TO CORPUS INVESTIGATION

Claire Kennedy and Tiziana Miceli
Griffith University, Brisbane

ABSTRACT

This paper reports on our experience in using a corpus of our own compilation, *Contemporary Written Italian Corpus (CWIC),* in teaching intermediate students at Griffith University in Australia. After an overview of the corpus design and the training approach adopted, we focus on our initial evaluation of the effectiveness of the students' investigations.

Much has been written on *what* can be done with corpora in language learning: what kinds of discoveries can be made with different types of corpora. There is relatively little on *how* learners actually go about investigations. Since we intend for our students to progress from classroom use to independent work as a result of using a Web-based version of *CWIC*, we have been seeking to understand how successful they are at extracting information from this corpus in the absence of a teacher. Our initial study highlighted the complexity of the process and the specialized skills required. We found that lack of rigor in observation and reasoning contributed greatly to the problems that arose, as did ignorance of common pitfalls and techniques for avoiding them. We, therefore, conclude the paper with an outline of proposed changes to our apprenticeship program, aimed at better equipping the students as "corpus researchers."

INTRODUCTION

Much of the literature on the use of corpora in language teaching relates to courses for advanced and highly motivated students of English for Specific/Academic Purposes (e.g., Johns, 1988, 1991a, 1991b; Levy, 1992; Mparutsa, Love, & Morrison, 1991; Stevens, 1991; Tribble, 1991) or translation (e.g., Aston, Gavioli, & Zanettin, 1998; Bernardini, 1998; Gavioli, 1996). So, when contemplating the introduction of work with corpora into the undergraduate Italian program at Griffith University in Australia, we were aware of the need to tailor the experience to quite a different target group, for whom Italian is usually a foreign rather than a second language and whose intentions for its use are less ambitious. In other contexts, our students might be regarded as reaching intermediate or higher-intermediate levels.

Our aim was to provide these students with a corpus to use primarily as a reference resource while writing. In view of their proficiency level and the types of writing tasks in which they engage, we sought a corpus that would supply models of personal writing on everyday topics. At the time, the only corpus of contemporary written Italian available to us was a collection of newspaper material,[1] so we first resolved to create our own corpus, which we have named *CWIC*, or *Contemporary Written Italian Corpus.* Secondly, we also decided to initiate the students into corpus use in a gradual and guided manner and thirdly to attempt to evaluate the effectiveness of their work with *CWIC* as soon as possible. This paper discusses the implementation of those three decisions, with the focus on an initial evaluation exercise and its implications for our approach to training students.

The *CWIC* Project: A Corpus for our Teaching and Learning Context[2]

Most of our students begin their university experience with no prior knowledge of Italian and can attend a maximum of 400 contact hours during their three years in our program. Additionally, they are not usually able to spend time in Italy during their studies nor are there many local opportunities for immersion. We estimate that, on the average, they graduate with "basic vocational proficiency" in reading and listening, on the scale used in Australia (Wylie & Ingram, 1999), while their ratings in speaking and writing are lower, somewhere between "basic social proficiency" and "basic vocational proficiency."

In their second year, the students begin intensive writing practice with letters and diaries, some creative writing and informative pieces based on their own experience. In their third year, the work is more academic in the sense that they bring their analytical skills to bear on the topics. The writing tasks are defined as commentaries, reviews or short essays, and treat aspects of the novels and films studied or television news items and newspaper articles.

In designing *CWIC* for this context, we were informed by various reports on the merits of small corpora for language learners, especially Tribble's advice that "the most useful corpus for learners … is the one which offers a collection of expert performances in genres which have relevance to the needs and interests of the learners" (1997, p. 3) and Aston's recommendation of corpora restricted to familiar text types and topics (1997, p. 62). We envisaged *CWIC* as complementing the newspaper corpus by providing models of texts by non-professional writers, including personal correspondence, although we chose to include some journalistic writing as well. We refined our general selection criterion of contemporary written usage to the following: short, written texts of specific text types (see Table 1), produced since 1990, by adult native speakers of Italian using non-specialist language.

Table 1. Text types included in *CWIC*

By non-professional writers	private letters business and official letters private email messages business and official email messages email messages to mailing lists letters to experts in magazine columns
By professional writers	experts' responses articles in regular magazine columns film reviews

Within the constraints of physical access to texts and the feasibility of obtaining permission to use them, our selection has been motivated by the desire to include a range of topics that our students might find interesting or relevant, in texts likely to be comprehensible. The email lists and magazine columns are a valuable source of material on a wide range of themes.[3] Our interest in content stems from the expectation that students will come to appreciate the corpus not only as raw material for concordances and frequency lists but also as a database of whole texts, which can be interesting to browse through collectively or read individually.

At the time of writing this article, we have approximately 570,000 words, in 2,200 texts by 930 different authors.[4] While we make no claims regarding representativeness of language in general, we can say that *CWIC* provides models of expert performances in several of the text types that our students encounter and are required to produce, during as well as after their studies. It also offers a wealth of appropriate language that can be used in other writing tasks such as creative writing and essays.

The Students' Apprenticeship

Since Johns (1988, p. 24) raised the need for learners to develop strategies of observation for extracting information from the data, many teachers experimenting with concordancing in the classroom have favored a gradual and guided approach (Johns, 1991b, p. 31; Stevens, 1991, p. 39; Tribble & Jones, 1997 p. 58). Guiding learners through a series of preliminary concordance-based activities has been presented as a way of both familiarizing them with various types of investigations that can be conducted and stimulating this development of appropriate learning strategies through practice (Turnbull & Burston, 1998, p. 18).

We too opted for an “apprenticeship” approach to training the students, intended to promote learning by example and by experience. We began with the second-year cohort, in a subject that includes a weekly two-hour writing workshop. For most of the training we used a sub-corpus of 50,000 words containing texts of each type, so the students could become familiar with the corpus without facing vast arrays of examples. The activities were initially carried out step by step, with the teacher giving directions through a series of leading questions, sometimes calling attention to particular examples. The students worked in pairs or small groups and reported back to the rest of the class.

Interrogation of the corpus was not presented as an end in itself but rather as an integral part of the writing and grammar work being undertaken. There is considerable attention to morphology and syntax in the subject, since it is at intermediate level. We started concordancing activities in that context, by examining verb constructions with direct and indirect objects as well as the behavior and meaning of certain conjunctions and pronouns.

After the first few sessions, we began to encourage the students to use the corpus while revising their own written work. Periodically, we presented the class with anonymous sample sentences from the previous week's writing and worked with them on ways of using the corpus to make corrections. In this way, they practiced formulating questions, such as "Should we use *infine* or *finalmente* here?", and devising appropriate searches. When marking their work, we pointed out where they might be able to make corrections themselves, with reference to the corpus. This meant dedicating some class time to individual problem-solving work, with the teacher circulating to assist as needed.

Finally, we presented applications of the corpus in composing and in pre-writing work, for what we call "treasure-hunting": finding models of ways to express things. Several such activities were conducted with a sub-corpus of personal letters. The students first browsed freely through several letters, observing typical opening and closing sequences. Then, they looked for ways of expressing certain functions, such as apologizing for not writing sooner, thanking someone for a previous letter, or giving information on chosen topics such as work, family, or exams. They did this both by skimming sequentially and by searching on words they thought might be present. For example, *ricevere* produced the expression *Non sai che piacere mi ha fatto ricevere tue notizie* (You don't know how pleased I was to receive your news) and *vita* turned up *La mia vita sentimentale è veramente uno schifo* (My love life is truly lousy). The students also examined frequency lists for combinations of three or four words, which brought to light a host of useful sequences, such as *Non vedo l'ora* (I can't wait), *Ci sono novità?* (What's new?) and *al più presto* (as soon as possible). These proved to be interesting and entertaining to the students, not only as alternatives to overused expressions, but also as triggers for further searches.

Neither in problem-solving nor in treasure-hunting work did we seek to engage the students in free exploration without a predetermined aim. There was always a defined goal: to find out how to phrase something specific in a given text. However, some experimented with "serendipity learning" (Johns, 1988, p. 21) during treasure-hunting activities and we encouraged them to continue to do so in their own time.

Overall, we viewed this introductory semester as a time for preparing students for independent mode future work with larger corpora outside the classroom. Until now, corpus interrogation has been performed using the text database software *DBT3 Database Testuale* (Picchi, 1997), which is installed in our laboratories. We believe that *DBT* has a friendly and intuitive user interface and that offers an appropriate range of functions, including concordancing on single words or expressions (which can be quite complex), labeling of examples to identify sources, and sorting. Moreover, the length of context displayed for each example is configurable, and clicking on an example expands the context to full-screen. There is also a browser for viewing whole texts and the battery of reporting functions includes frequency lists. Soon, however, students will have access to *CWIC* from home, since we are currently working to transfer it to a Web platform, with its own searching software, offering functionality and tools similar to *DBT*. In 2001, we will be involving our students in a pilot exercise using *CWIC on the Web*.

A total of 7 class contact hours out of a total of 26 hours in the writing workshop strand of the subject were dedicated to *CWIC* during the semester. The students also worked with the corpus on assignment tasks for a few hours outside of class time. Only 3 of the 17 students who completed course evaluation questionnaires said that the amount of time spent was disproportionate to the usefulness of the exercise. In the future, we intend to more closely examine the relationship between the time invested in concordancing work training and the benefits attributable to the mastery of this type of reference tool.

The questionnaires, combined with class discussions and individual interviews, were intended to draw out students' perceptions of certain aspects of the corpus induction experience. Because these findings are the subject of a separate study, we will only mention some of the main points here. On the positive side, most students reported that work with the corpus helped them to better understand Italian grammatical structure and boosted their confidence in correcting their own writing. Their various definitions of what made the corpus a useful resource can be grouped into three categories: it provides examples of real language; it allows exploration of the various uses of a given word in different contexts; and it illustrates the specific functions of certain words and expressions in particular types of text. On the negative side, some stressed the discouragement felt on not being able to understand all the examples or to identify relevant ones, and most admitted that they had on occasion found searches too time consuming and frustrating. Our first evaluation exercise was concerned with this aspect: what creates a successful investigation and what causes unproductive searches and frustration.

EVALUATION: AIMS AND PROCEDURE

In view of the proficiency level of our students and our intention that they use *CWIC* and other corpora outside the classroom, we were keen to understand how effectively they were able to use it on their own, specifically the mechanics of their investigations and the difficulties they encountered. We found little to inform us in developing an approach to such a study. Flowerdew (1996, p. 112) drew attention to "a paucity of critical perspectives in concordancing literature," but his call for more in-depth evaluative work does not appear to have borne fruit. Much has been written on *what* can be done with corpora in language learning -- what kinds of investigations can be conducted with different types of corpora and what kinds of discoveries are made, usually in a classroom context -- but relatively little in the literature on *how* students actually do this, and especially on how they fare on their own.

Two of the studies we located, however, do reflect an interest in evaluating students' independent work. Turnbull and Burston (1998) analyzed the aims and outcomes of investigations conducted by advanced students after only minimal training with a concordancer, but mainly with the goal of demonstrating the importance of adequate training. Closer to our purposes was Bernardini's (1998) examination of the processes and outcomes of students' exploration of the British National Corpus, as a result of which she outlined suggestions for making this kind of work more systematic. Among the tendencies she noted were ignoring variants, not looking for alternative approaches when faced with an obstacle, and making only a summary analysis. While our students were not involved in free exploration of a large corpus, we anticipated that these kinds of problems were likely to characterize their work, too.

We chose to focus on our students' handling of problem-solving activities while revising a text as the first stage of our evaluation since much of their work done in class had been like this. Essentially by asking them to "show their work," we collected data on how they went about using concordances to answer specific questions while correcting their own or others' work.

The 10 students referred to in the discussion that follows (S1 to S10) came from the top and middle ranks of the cohort in terms of their achievement in our subjects. For the purposes of this paper, we numbered them according to their results in the written Italian subject in which the corpus apprenticeship was conducted as well as in its companion subject in spoken Italian. S1 was the top performer. Of the 10 students, 5 were enrolled in a languages and linguistics degree program, and the other 5 were studying history, law, or psychology.

Some of the cases cited are drawn from activities individually carried out by the students during the semester. Evidence comes from their own accounts of how they used *CWIC*, sometimes in tasks set by us but oftentimes in those they set for themselves while in the process of editing their own compositions on given topics.

The majority of the cases come from pair-work sessions held immediately after the end of semester, which were video-recorded and followed immediately by an interview aimed at extracting a retrospective account of the students' work. Eight students participated, and the sole criterion for pairing them was their availability at particular times. They were given two texts to revise. In the first, we set specific tasks by underlining certain words to indicate where there might be a problem. In the second, we invited them to decide what issues to deal with for themselves.

We expected that in the investigations they initiated themselves the students would work on relatively familiar language points, approached with some degree of confidence. The set tasks, on the other hand, were intended to force them to address types of problems they might not otherwise tackle.

In both the individual and pair-work situations, all the texts we provided for the activities had been selected from work submitted by students in that subject. Dictionaries and grammar books as well as the corpus were available at all times. The students were encouraged to use all three resources as they deemed appropriate.

RESULTS

Overview

We found that the students made many successful investigations, demonstrating a general appreciation of the types of questions that can be posed, a certain ability to work by analogy, and a preparedness to review their strategies when a search was leading nowhere. However, our concern was to identify what went wrong or could be done more efficiently, in order to gain insight into how to improve the apprenticeship. Our observations suggested that, while knowledge and experience of the language undoubtedly played a part in how productive the students' work with the corpus was, lack of rigor in observation and reasoning contributed greatly to their difficulties, as did apparent ignorance of common pitfalls and techniques for avoiding them. We concluded that our training had not adequately equipped them as "corpus researchers."

Our Analysis of Learner Investigations

In order to understand what happens in a corpus investigation, we approached it as a four-step process: (a) formulating the question; (b) devising a search strategy; (c) observing the examples found and selecting relevant ones; and (d) drawing conclusions. This schema is illustrated below with reference to one of the set tasks from the pair-work sessions. The sentence concerned was *Sto cercando l'orario per il corso LAL3093* (I'm looking for the timetable for the subject LAL3093). The problem is the choice of preposition: *per* is often used where *for* is used in English, but not in this context, where the pattern is *orario di.* An appropriate and efficient way of dealing with the task is described in Table 2.

Table 2. Steps in a Corpus Investigation

1.	Formulate the question	"Which preposition can be used after *orario* when speaking of a *timetable for something?"*
2.	Devise a search strategy	Search on *orario,* with a view to checking what follows it.
3.	Observe the examples and select relevant ones	Look for examples in which the idea *timetable for something* is expressed.
4.	Draw conclusions	Check which word(s) are used with *orario* in those examples. Identify the combination *orario di* and insert it into the target sentence, making any necessary adaptations.

Occasionally, the students' investigations did not conform exactly to this pattern, as they had no clear question in mind at the outset of their search. This happened if they were working on a set task and had no idea what the issue might be, so they performed a preliminary search on the underlined word or neighboring ones, just to see what came up. If nothing attracted their attention, they abandoned the task, but if they did notice something they formulated a question and then proceeded through the remaining steps as outlined above.

The discussion that follows examines students' work on each step in some detail. We frequently use specific cases to illustrate the types of problems that led to an unsuccessful outcome. Despite this focus on what goes wrong, our intention is to convey how complex a corpus investigation is, rather than to present the students' performance as unsatisfactory. We trust the analysis serves to highlight the specialized skills the learners employed and the variety of factors they are required to bear in mind.

We have not included cases in which an unsuccessful outcome was caused by lack of linguistic knowledge, although we recognize that proficiency is important, especially in Step 1 and Step 3. In Step 1, for instance, appreciation of whether it makes sense to ask a given question depends to some extent on familiarity with the target language. In Step 3, of course, not understanding the examples can undermine even an impeccably conducted investigation. However, our interest here is in identifying problems that did not appear to result from inadequate proficiency and that could perhaps be overcome by appropriate training. We, therefore, sum up the discussion of each step in the form of a list of tips for learners. We do not present these as rules ready to be imparted to future groups of trainees, but envisage drawing them up together with the students, through collective reflection on investigations carried out in class.

Step 1: Formulating the Question

Before examining what goes on in this step, it is important to note what types of questions were being dealt with in the investigations. They were not free exploration questions such as "What can I find out about *x?"* nor treasure-hunting questions like "In what ways can I express this function?" Instead, the questions were aimed at checking or correcting a given sentence. Those we encountered in the students' work were of just three types: (a) "What is/are the correct word(s) in this context to render this meaning?"; (b) "What construction do I need around this word (or these words) in this context?"; and (c) "What order should these words be in, in this context?" Each type can be expressed in open form, as above, or in closed form. For example, two closed forms of the first type of question are: "Can *x* be used to render this meaning in this context?" (yes/no form) and "Is *x* or *y* the correct word for this meaning in this context?" (multiple choice). Clearly, we use the words "yes/no" here as shorthand for "there is / is not evidence for this."

We found that some of the questions the students formulated suggested that they might have misconceptions about the types of questions it is logical to ask, the kinds of information that can be obtained from a corpus, and the ways clusters of words behave. We grouped the problems we identified into five categories.

First, there was sometimes insufficient attention to how specific or general a question should be. When dealing with *Sto cercando l'orario* ***per*** *il corso,* S3 and S6 asked "Can you say *per il corso?"* This is too general a question; while the answer is "yes," this does not help decide whether these words can be used in the given sentence. It seemed that the students needed to be more conscious of the fact that the actual combinations of words used in a language are only a subset of the potential combinations (Gavioli 1996, p. 124). Reflecting on the implications of general or specific questions in their native language could help students appreciate this problem (provided they do not assume answers can be transposed). For example, "Can you say *for the course?"* is not a useful question in English either. We say *prerequisites for the course* but *aims of the course.*

An unnecessarily general question may well eventually lead to a successful outcome, but the investigation is likely to be inefficient, due to detours to deal with evidence in contexts not relevant to the case at hand. We observed this in students' handling of one of the individual set tasks, that of choosing between ***Il lunedì scorso*** *siamo andati all'università* and ***Lunedì scorso*** *siamo andati all'università* for *Last Monday we went to university.* The issue is whether the definite article is used with *lunedì scorso* (last Monday) and the answer is "no." Some asked the question "How does *scorso* (last) behave?" rather than "How does *lunedì scorso* behave?" This meant dealing with *scorso* in several contexts, some with an article and some without.

Second, the students often did not seem to consciously choose whether to frame their questions in open or closed form. Primarily, they did not take into consideration that a closed question could lead them to a dead end and the need for a follow-up question. This happened to S1 and S7: after dealing with their question "Do you say *orario per?"* they found that they needed a second investigation aimed at answering "So what *do* you use after *orario?"*

The third type of problem was apparent when a question arose only after looking at some examples. In this situation, students sometimes failed to formulate the question explicitly. One of the set tasks in the pair work was to check the sentence ***Auguri*** *per il weekend,* with which the writer had intended to say something like *Have a good weekend.* Here, *auguri* is out of place: it usually corresponds more to *best wishes* and is used for birthdays and other special occasions. S10 had an idea along those lines, suggesting, "Maybe they don't say *wishes* for the weekend, maybe they mean *wishes* like *congratulations,"* but she and S8 did not turn that into the question, "So how *do* you say *Have a good weekend?"* which might have led them to search on other words. They just continued to muse upon the examples of *auguri* and eventually gave up.

Fourth, there was the fatal lure of prepositions. The students' attention was often attracted to a preposition itself rather than to the words around it, on which it depended. In some cases, they treated a preposition as having a meaning in isolation, or as being in one-to-one correspondence with an English preposition, such as when S4 said to S9 "Doesn't *da* usually mean *from?"* Very common indeed was the habit of treating a preposition as linked only to the words following it. For example, when correcting her own sentence *Il cane è troppo stanco … continuare il gioco* (The dog is too tired to continue the game), S5 asked "What preposition do I want before *continuare?"* rather than "How do I construct *too <adjective> to do something?"*

Fifth, there was a tendency to neglect lexical considerations in favor of grammatical ones, to focus on how to combine words rather than whether they could be used at all in the given context. For example, when presented with *Non mi sorprenderebbe* ***imparare*** *che ho fatto molti errori* (It wouldn't surprise me to learn I've made many mistakes), nearly all the students considered only the construction of the sentence. They did not question whether *imparare* could be used for *learn* in this sense.

On the basis of our observations, an initial set of tips for Step 1 might be as shown in Table 3.

Table 3. Examples of Tips: Step 1

- Try to state your question precisely.
- Ensure it is specific enough for the situation you are dealing with.
- If it is in yes/no or multiple choice form, consider whether an open question would be more appropriate. For example, rather than asking "Does *y* come after *x*?" you might want to ask "What comes after *x*?"
- Keep in mind both lexical and grammatical issues.
- In your dealings with prepositions:

  When considering what word(s) a preposition might be linked to, look both to the right and to the left, and to a distance of a few words.

  If you are trying to choose a preposition for a particular context, remember the possibility that *no* preposition is required there.

Step 2: Devising a Search Strategy for a Given Question

We identified the components in the definition of a strategy as (a) choosing the word(s) to search on and (b) deciding whether and how to use other options such as sorting examples or consulting a dictionary or grammar book. Choosing the word(s) to search on is not necessarily just a matter of deciding which are the key words in the question. It may entail picking words that can be substituted for these, such as different forms of a lemma or words that belong to the same set (like days of the week, colors, possessive pronouns).

Students did not always pay sufficient attention to exactly defining the construction they were dealing with and therefore distinguishing its fixed and variable parts. Often this coincided with a certain difficulty in framing the question. One example of many was the treatment of *Non mi sorprenderebbe imparare* (It wouldn't surprise me to learn) by S1 and S7. They wondered whether a preposition is required between the conjugated verb *sorprenderebbe* and the infinitive *imparare.* Their strategy was to search on *imparare.* It did not seem to occur to them that it was the behavior of *sorprendere* that mattered, that the construction is a variant of *Non mi sorprenderebbe <infinitive>* or, more generally, *Non <object pronoun> <conjugated form of sorprendere> <infinitive>.*

Nor did students seem very concerned that a strategy be efficient. That is, they did not direct effort at obtaining a workable number of examples -- not too many -- with as many as possible of them likely to be relevant to the problem at hand. This means including as much as possible in a search combination without, of course, prejudicing a successful result by making it too restrictive. During the pair work, S4 and S9 set themselves the task of deciding between *niente da fare* and *niente di fare* for *nothing to do.* They searched on *fare* (to do) and sorted the examples so as to check on the left for *di fare* and *da fare.* Since *fare* is present in a myriad of idiomatic expressions, it provided a host of irrelevant examples to sort and scroll through.

Additionally, there were several cases of students overlooking the option of trying other forms of the key word (or substituting another word for it altogether) in the event of not finding any examples. When dealing with Italian nouns, verbs, and adjectives, if a first search fails to produce sufficient evidence, it should be automatic to check different inflected forms. In one instance, S1 posed the question of whether the adjective *estrema* (extreme, singular feminine) should precede or follow the noun and made her decision on the basis of only one example. Had she searched on the masculine and plural forms as well (or used the stem together with a wildcard character: *estrem*),* she would have had several examples to consider, and she would have been able to detect the mobility of this adjective, with the choice of position reflecting degree of emphasis. Another aspect of this issue is that if students *did* think to search on another form of a verb, they tended to only try the infinitive, apparently transferring dictionary practice to corpus use.

Clearly, there are many factors to take into account when devising a strategy, and it is not surprising that the students did not always think of all possible ways of fine-tuning their approaches. There were several occasions in which they neglected to use certain options to their best advantage. For example, when searching for a combination of words, students sometimes forgot to specify if they were interested in the words only when they were adjacent. This is quite simply achieved by setting a maximum-distance-apart parameter to 1. We noticed the converse problem too, of setting this parameter to 1 automatically, without considering whether the search words were likely to be separated in the examples by intervening words, phrases or even clauses. Sorting features were also used somewhat indiscriminately. The words linked to a keyword may well not be adjacent to it, and looking at sorted output sometimes distracted the students' attention from useful examples.

Finally, there were times when the students were apparently so engrossed in the corpus that they forgot to use the dictionary. This was noticeable at moments when they realized they were dealing with a word that did not have the desired meaning in a certain context. They got as far as checking the wrong word in the corpus and establishing that there was no evidence for using it in the target sentence, but then simply relied on their own memory or imagination in determining what to use instead, rather than reaching for the English-Italian dictionary.

In light of these observations, we drew up a basic set of tips for Step 2, shown in Table 4.

Table 4. Examples of Tips: Step 2

- Think about how efficient your strategy will be. Is it likely to generate many irrelevant examples alongside the useful ones? If so, maybe you should restrict your search further.
- Check if you are dealing with a variant of a general pattern, with a fixed part and a variable part, as you may want to search only on the fixed part.
- If you are not satisfied with the examples found, think about using wildcards or substituting something else for one of the search words: another form of the same lemma or a word that may be equivalent in the context that interests you.
- Remember the English-Italian dictionary if you are looking for potentially appropriate words.

Step 3: Observing the Data and Selecting Examples

Surprisingly often, students lost sight of the importance of selecting examples with a view to matching form and meaning closely to the requirements of the target sentence. For example, while editing a sentence using the adverb *ancora* to mean "again," S8 investigated the behavior of *ancora,* saying that she was interested in its position with respect to the verb it modified. Her eventual construction was fine, but in none of the four examples she cited as her models did *ancora* have the meaning *again* nor did it always modify a verb.

Most of the time, the students did check the meaning and structure of examples, but they were not always bent on finding a close match, even when excellent evidence was readily available. It became clear to us that there were specific traps for those who were anything less than rigorous in the selection of examples. One was the distraction offered by a majority of examples being of one kind. In this situation, useful examples belonging to a minority category were easily ignored. In the case of *lunedì scorso,* most students who searched on *scorso* were attracted by the examples of *il mese scorso* (last month) and *l'anno scorso* (last year), which include the definite article, and used these as their model. This was despite the fact that 2 of the 15 examples found illustrated exactly what was needed, in the form of *last Monday* and *last Thursday,* without the article. Another trap for students who did not attend closely to meaning was the way some combinations turn up due to the chance juxtaposition of phrases, not because they form a lexical phrase themselves.

A frequent problem was that of students not noticing something if it was not what they were looking for. This was the case when S4 and S9 were trying to establish whether they should use *cercare a* for *to look for.* They simply did not see the several examples on the screen of *cercare* used transitively to mean *to look for*, because they were intent on choosing a preposition. The problem of not noticing all the information given could be observed also at the moment of applying an example as a model in the target sentence. In an individual task, S9 wanted to see what verb construction to use in *After bringing the stick back,* and her first attempt included *dopo restituendo* (after <gerund>). She then looked up *dopo* and found an example, which included *dopo aver chiuso* (after having closed), or *dopo aver <past participle>*. However, she appeared to notice only the pattern *dopo aver,* and so she just inserted *aver* into her first guess, producing the hybrid *dopo aver restituendo.* Some tips for step 3 are shown in Table 5.

Table 5. Examples of Tips: Step 3

- Remember to check the meaning of examples you want to use as evidence, and seek out those that most closely match the requirements of your target sentence.
- Try not to be influenced by assumptions about what you will see in the examples. Look to the left and right of keywords to see which words are linked to them. The words you are expecting to find may not be present, and vice versa.
- Try not to be attracted only to the types of usage of a word that occur most frequently. The type you are interested in may be a less common case.

Step 4: Drawing Conclusions

The observations we made on this step primarily concern problems in reasoning, particularly the implications students drew from the number of examples found by a search. When only one or very few examples were found, the students tended to lack confidence in the result, evidently assuming that many illustrative examples are necessary to establish a case. This reflected a lack of appreciation for the fact that what matters is the quality rather than the quantity of examples. Depending on the type of question addressed, one example that is suitably analogous to the target can be sufficient for a "watertight case."

On the other hand, of course, if only a few examples are found, they may be the result of chance juxtapositions, the reality being that no relevant examples are present. This suggests a more general issue about numbers of examples: If many turn up when few are expected, or vice versa, the significance of this should be considered. The students sometimes expressed perplexity (for example, S10 said at one point, "Wouldn't you think there'd be a lot more examples?") but failed to act on this dilemma.

Various invalid conclusions were drawn at times when no examples were found. These included, "The phenomenon does not exist" in place of "There is no evidence for it in this corpus"; "The answer is not *x* so it must be *y*" when it was only a matter of supposition that *x* and *y* should be the only options; and "The search didn't work" or "We didn't find out anything," because the search had not produced the expected results. Some tips based on these observations are shown in Table 6.

Table 6. Examples of Tips: Step 4

- Even if you have only one example as evidence, it may be enough on which to base your case. Remember that what matters is how good your evidence is, not how much of it there is.
- If you have found only a few examples when you were expecting many, or vice versa, you may need to think about what this means. Why were you expecting to find many or only a few? What has affected the result?
- If you have found no examples, think carefully about what conclusion you can draw. Make sure you relate your conclusion to the question that you initially posed.

WHERE TO FROM HERE?

In the investigations we analyzed, difficulty in understanding examples was very rarely the sole or even the primary cause of invalid results. In fact, the above discussion is based entirely on cases in which it was unlikely that the examples dealt with were hard for the students to understand. Furthermore, in the pair work it was often the student with lower proficiency (as far as that can be measured by results in our subjects) who appeared more competent in using the corpus to tackle a problem. There were many instances in the pair work where S7, S9 and S10 led the way for S1, S4, and S8 respectively, by showing insight in formulating a question, using clear reasoning in devising a strategy, or paying attention to examples.

By this we do not mean to suggest that language proficiency is irrelevant nor to deny how daunting arrays of examples can be. We simply intend to underline that, in each of the four steps, we identified specific problems that seemed to be due to inadequate corpus-investigation skills. These were accompanied by an evident lack of awareness on the students' part of how easy it is for an investigation to be derailed.

So the apprenticeship now appears far more complex than we had thought. The evaluation has highlighted the need to focus on treating the students as trainee researchers. As in any other field of research, it is necessary for novices to acquire certain attitudes and habits of reasoning. They need to become acquainted with underlying principles and to master specific techniques, which are not necessarily intuitive. We are, therefore, reviewing our approach in two main areas.

First, we are looking for ways to encourage students to distinguish between observation and interpretation of data so as to try to free the observation phase of assumptions. To prepare the students to exploit the "direct access to the data," that a corpus provides (Johns, 1991b p. 30), we must convey to them the importance of observation rigor to precede interpretation of what is observed. This means that work aimed at raising consciousness of the idea that language is made up of "lexical phrases" rather than single words and that putting a sentence together is a matter of arranging patterns of words, attending to "the ways they can be pieced together, along with the ways they vary and the situations in which they occur" (Nattinger, 1980, p. 341). Careful observation in relation to these aspects can be expected to help overcome assumptions.

In addition to including explicit observation exercises in the training program, we are also making a much more general change. In order to "market" the benefits of observation to the students, we have decided to entirely reverse the order of our approach so as to start with treasure-hunting, and borrowing chunks of appealing language while composing texts. Subsequently, we will move on to the use of concordances to solve specific problems regarding word use, while revising texts. In this way, we mean to highlight, from the outset, the value of a corpus as a database of whole texts and of models of complete utterances and set phrases. We hope that by beginning with treasure-hunting we can encourage students to appreciate exploration of the corpus without prior assumptions about the data that will be found and cultivate in them a more open mind towards the ways strings of words belong together.

The second key aspect of our new approach, as foreshadowed in the preceding section, will be that of engaging the students in reflection on the processes of their problem-solving investigations. Once they have some experience in using the corpus to answer questions, we will introduce exercises -- perhaps presented in the form of "spot what goes wrong" -- aimed at collectively deriving a checklist of tips along the lines of those drafted above.

CONCLUSION

We recognize that during corpus investigations by language learners, there is considerable room for error due to lack of knowledge of the target language. However, we propose that the development of appropriate research habits -- incorporating observation and logical reasoning as well as techniques in corpus searching -- could reduce other causes of error to a minimum. Although we do not go so far as to suggest that learners need formal training in logic in preparation for corpus work, our evaluation of the ways students go about problem-solving with *CWIC* has convinced us of the importance of an awareness of logical principles applicable to this kind of operation.

The plan outlined above to revise our approach to training reflects this conviction. We expect that an apprenticeship oriented toward the development of "corpus research" skills will not only help students make the most of corpora but will also benefit other areas of their language learning as well, enhancing their capabilities with other reference tools in particular. Our next step will be to examine the effectiveness or lack thereof of the new approach, especially in work with *CWIC on the Web.*

NOTES

1. The Corpus of Italian Newspapers, available from the Oxford Text Archive at http://ota.ahds.ac.uk, contains 1,200,000 words from four dailies.
2. A more detailed description of the corpus and compilation process is in a paper submitted for the proceedings of the conference Teaching and Language Corpora 2000.
3. Some of the themes of magazine columns selected so far are health, education, personal problems, young people's issues, pet care, home computing, current events, social issues, science, and spiritual and theological questions. We have explored email lists belonging to groups of women, gays and lesbians, animal liberationists, translators and interpreters, vegetarians, mountain climbers, Italians overseas and fans of Totò, and on issues to do with politics, entertainment, current events, and personal problems.
4. The composition of the corpus is roughly 50% email, 5% letters, 40% magazine material (including letters from the public), and 5% film reviews. Non-professional writers account for over 75% of the content. The number of texts by a single author ranges from 1-10 for most of these to 30-40 for magazine column hosts.

ACKNOWLEDGMENTS

We thank Dr. Mike Levy and the anonymous reviewers for valuable feedback on an earlier version of this paper.

CWIC was developed with the assistance of grants from Griffith University and the Australian Government's Committee for University Teaching and Staff Development.

ABOUT THE AUTHORS

Claire Kennedy and Tiziana Miceli are lecturers in Italian at Griffith University in Brisbane.

Email: C.Kennedy@mailbox.gu.edu.au, T.Miceli@mailbox.gu.edu.au

REFERENCES

Aston, G. (1997). Enriching the learning environment: Corpora in ELT. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles, (Eds.), *Teaching and language corpora* (pp.51-64). London: Longman.

Aston, G., Gavioli, L., & Zanettin, F. (Eds), (1998). *Proceedings of corpus use and learning to translate conference,* University of Bologna, Bertinoro. Retrieved November, 8, 2000, from the World Wide Web: http://www.sslmit.unibo.it/cultpaps.

Bernardini, S. (1998). Systematising serendipity: Proposals for large-corpora concordancing with language learners. *Proceedings of TALC98* (pp. 12-16). Oxford, UK: Seacourt Press.

Flowerdew, J. (1996). Concordancing in language learning. In M. Pennington (Ed.), *The power of CALL* (pp. 97-113). Houston, TX: Athelstan.

Gavioli, L. (1996). Corpus di testi e concordanze: Un nuovo strumento nella didattica delle lingue straniere [Text corpora and concordances: A new tool for foreign language teaching]. *Rassegna Italiana di Linguistica Applicata, 2,* 121-146.

Johns, T. (1988). Whence and whither classroom concordancing. In T. Bongaerts, P. De Haan, S. Lobbe, & H. Wekker (Eds.), *Computer applications in language learning* (pp. 9-27). Dordrecht, The Netherlands: Foris.

Johns, T. (1991a). Should you be persuaded: Two samples of data-driven learning materials. *English Language Research Journal, 4,* 1-16.

Johns, T. (1991b). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *English Language Research Journal, 4,* 27-45.

Levy, M. (1992). Integrating computer-assisted language learning into a writing course. *CAELL Journal, 3*(1), 17-27.

Mparutsa, C., Love, A., & Morrison, A. (1991). Bringing concord to the ESP classroom. *English Language Research Journal, 4,* 115-133.

Nattinger, J. (1980). A lexical phrase grammar for ESL. *TESOL Quarterly 14*(3), 337-344.

Picchi, E. (1997). *DBT3 Database Testuale.* Consiglio Nazionale delle Ricerche, Italy. Distributed by Lexis Progetti Editoriali s.r.l. See http://www.lexis.it.

Stevens, V. (1991). Classroom concordancing: Vocabulary materials derived from relevant, authentic text. *English for Special Purposes Journal, 10,* 35-46.

Tribble, C. (1991). Concordancing and an EAP writing program. *CAELL Journal, 1*(2), 10-15.

Tribble, C. (1997). *Improvising corpora for ELT: Quick-and-dirty ways of developing corpora for language teaching.* Paper presented at the first international conference "Practical Applications in Language Corpora," University of Lodz, Poland. Retrieved November 8, 2000, from the World Wide Web: http://ourworld.compuserve.com/homepages/Christopher_Tribble/Palc.htm#Top.

Tribble, C., & Jones, G. (1997). *Concordances in the classroom: Using corpora in language education.* Houston, TX: Athelstan.

Turnbull, J., & Burston, J. (1998). Towards independent concordance work for students: Lessons from a case study. *ON-CALL*, *12*(2), 10-21.

Wylie, E., & Ingram, D. (1999). *International second language proficiency ratings: Master general proficiency version (English examples)*. Brisbane, Australia: Centre for Applied Linguistics and Languages, Griffith University.