An apprenticeship with the CWIC corpus: a tool for learner writers in Italian

Tiziana Miceli and Claire Kennedy (Griffith University)

Abstract

This paper discusses the composition and application in teaching of the Contemporary written Italian corpus (CWIC), created at Griffith University as a resource for Italian students at intermediate level. Since 1998, CWIC has been used in our second-year and third-year courses, primarily as a reference tool while writing. Our rationale in introducing CWIC drew particularly on the idea that learners armed with a corpus can act as researchers who discover rules about language for themselves, by formulating and testing hypotheses with reference to authentic examples of word use in context, as outlined by Johns (1991b). We conducted a gradual and guided apprenticeship in corpus use over a whole semester, intended to prepare the students sufficiently to work independently with CWIC when a web-based version comes online. We then sought to evaluate the effectiveness of the students’ investigations and identify the difficulties they encountered when working without a teacher.

Although the students were very enthusiastic about using CWIC, our analysis of their investigations showed they were less successful than we had anticipated. In particular, it highlighted two types of problems – in managing a hypothesis-testing process and in observation of data – that significantly reduced their effectiveness in the ‘learner as researcher’ role. We are seeking to address these difficulties by changing the objectives and priorities of our training. A key element in our response is to introduce students to CWIC with an ‘observe and borrow’ mentality first, before progressing to an ‘observe and derive rules’ approach such as we had originally envisaged would be the norm.

1 Introduction

In the world of corpora and language learning, ‘classroom concordancing’ was the novelty of the 1990s, as improvements in computers’ speed and capacity meant that students and teachers could have hands-on access in laboratories to tools for producing concordances in real time, rather than working only with printouts generated beforehand by a teacher using a large computer. For the 2000s, however, further technological advances make it possible to go beyond just classroom use and beyond just concordancing. First, web-based corpora with their own search engines can allow individual access from anywhere, with users no longer needing a concordancing program on their own computer. Second, new types of software have been produced that treat a corpus as a database of whole texts, available for browsing and copying from as well as for producing concordances.

Our decision to introduce work with corpora into the Italian Studies programme at Griffith University stemmed from our perception that the expanding accessibility and functionality of corpora and related tools made the ability to use them an ever more valuable asset for language learners, both during and beyond their formal studies.
However, most of the literature on the use of corpora by language learners referred to students with apparently much higher proficiency in the target language than ours. Many of the students concerned were enrolled in advanced courses in English as a second language (ESL) or English for special purposes (ESP), often in preparation for tertiary or postgraduate studies in an English-speaking country (e.g., Johns 1988, 1991a, 1991b; Levy 1992; Mparutsa, Love & Morrison 1991; Stevens 1991; Tribble 1991), or degree programmes in translation and interpreting (e.g., Aston, Gavioli & Zanettin 1998; Bernardini 1998; Gavioli 1996). We perceived the need for particular attention to the choice of a corpus suitable for our, less advanced, students and the development of a training approach that would properly equip them for effective, independent use of it.

This paper first describes the corpus we have compiled, CWIC or Contemporary written Italian apprenticeship corpus, and outlines the approach we initially took to the students’ apprenticeship, devised with reference to the reported experiences of others and our understanding of relevant pedagogical principles and assumptions. It then discusses the way in which reflection on our practice and evaluation of its outcomes has led us to a revision of that approach and a changed appreciation of what work with corpora can offer.

2 Our teaching context and CWIC

At Griffith University we adopt a communicative approach to teaching Italian, but with considerable attention to grammar as essential to the construction of meaning. Corpora therefore appealed to us as a potential means to facilitate the students’ own exploration of grammar through observation and analysis of authentic texts. We decided to introduce a corpus to our students primarily as a writer’s reference resource, as we thought its usefulness would be most apparent while writing, it being a non-spontaneous process that allows time for reflection.

Corpora used for language-learning purposes tend to be relatively small, usually with fewer than one million words, unlike those developed for linguistic research. Concern for the comprehensibility of language and familiarity of text types and topics outweighs that for representativeness (Aston 1997; Tribble 1997). In seeking a suitable corpus of written Italian for our context we were guided in particular by Tribble’s recommendation 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” (1997:3).

Most of our students have no prior knowledge of Italian on entering our programme, which offers them a maximum of 400 contact hours, so we estimate that, on graduation, they reach what might be considered intermediate or higher intermediate level in other contexts. Much of their writing is of a personal kind – letters and emails, diaries, informative pieces based on their own experience – or creative writing. In third year they are also set more academic tasks, to produce commentaries, reviews or short essays treating aspects of novels and films studied or news and current affairs material. The writing they do for their own purposes, outside the course requirements or after completing their studies, may be of various types but is likely to include private correspondence.
We therefore sought a corpus that would provide abundant models of personal writing on ‘everyday’ topics, as well as texts by professional writers such as journalists and film critics. At the time, the only corpus of contemporary written Italian available to us was a collection of newspaper material, so we resolved to build our own, which we have named CWIC, or Contemporary written Italian corpus. In compiling it, our general criteria were that texts be short and produced after 1990 by adult native speakers using non-specialist language. Within this framework, we chose the text types listed in Table 1, in view of their ready availability and the likelihood they would supply material on a range of topics. The selection of individual texts hinged on our – obviously subjective – assessment of whether they were likely to be comprehensible, interesting and relevant to our students. So far we have about 570,000 words, in 2,200 texts by 930 different authors. A more detailed description of the corpus and the rationale behind the compilation process is given in Kennedy and Miceli (2002).

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

Table 1  Text types included in CWIC

3  Our initial approach to using CWIC

We trialled CWIC with second-year students in 1998 and 1999, in a second-semester subject that includes a weekly two-hour writing workshop. In this strand we seek to encourage students to appreciate the benefits – to their language learning in general – of treating writing as a process, involving pre-writing, composing and revising phases (Richards 1990:108-115). As this is the stage in our programme when students begin regularly producing their own texts, it seemed the appropriate point for presenting CWIC as a reference tool.

Since Johns raised the need for learners to develop strategies of observation for extracting information from concordance data (1988:24), many teachers have favoured a gradual and guided approach to introducing work with corpora in language classrooms (Johns 1991b:31; Stevens 1991:39; Tribble & Jones 1997:38). Taking the learners through a series of preliminary concordance-based activities has been presented as a way of both familiarising them with various types of investigation that can be conducted and stimulating the development of appropriate learning strategies through practice (Turnbull & Burston 1998:18). We followed this lead and opted for a kind of apprenticeship, spread over several months, intended to promote learning by example and by experience.
However, our approach to this apprenticeship differed in two main ways from those of the early 1990s that we drew on, as a result of the technological changes mentioned above. First, we envisaged classroom work not only as an end in itself but explicitly as preparation for self-directed work for individual purposes, outside class. Although our students have so far had access to CWIC only in the university’s laboratories, we have begun the operation of transferring it to a web platform with its own search engine, so as to make it accessible from the students’ desks at home, just like their dictionaries (whether printed or on-line). Second, we used a “text database management” package called DBT3 Database Testuale (Picchi 1997; see http://www.ilc.pi.cnr.it/PiSystem/), which not only provided searching (that is, concordancing) functions but also allowed browsing of whole texts, swapping between searching and browsing windows (especially between a concordance example and the complete text it comes from) and copying of text from a concordance or browsing window for pasting in other applications. We sought to exploit these features in our approach, so as to encourage the students to see the corpus not only as the raw material for concordances and frequency lists but also as a database of whole texts which can be interesting to browse through, read closely or copy from. Mike Scott’s Wordsmith Tools (1997) (http://www.lexically.net/wordsmith/) offered similar functionality, as will our custom-built CWIC-on-the-web facility.

We presented the students with two distinct modes of using CWIC: one concerned with correcting the language used in a text and the other with enriching its content and vocabulary. The latter we saw as particularly relevant to pre-writing and composing stages and the former to revising, but not exclusively.

3.1 Text-correcting mode

Most of the apprenticeship was dedicated to text-correcting work, or using concordances to improve the grammatical and lexical accuracy of a text. It is a problem-solving activity, in that the learner seeks answers to specific questions, usually to do with choosing the correct word(s) for a given context or assembling them in the correct structure or order. Some of the many types of questions that can be dealt with are: differences between near synonyms or false friends; verbal constructions (choice of auxiliary and preposition, overall structure); the position of adjectives; situations in which definite articles are required; and the meanings and structure of idiomatic expressions.

Our reason for focusing on this type of work was that we considered it more complex and challenging than text-enriching work (described below). There was also a wealth of ideas to draw on: particularly useful examples of types and techniques of corpus investigation are to be found in Concordances in the classroom by Tribble and Jones (1997) and the Micro-Concord Manual by Murison-Bowie (1993).

The immediate, practical aim of this part of the apprenticeship was to guide the students towards independent use of the corpus in the revision of their own written work. However, we recognised that, while self-correction is clearly an important overall goal of language learning, the extent to which it can be self-initiated is questionable. Often the existence of an error must be pointed out before a learner can self-correct, however highly-skilled with reference resources (Todd 2001:4). So we did not assume that equipping our students with the corpus would be a sufficient condition for them to
engage in systematic self-initiated self-repair. Nonetheless, by focusing the students’ work with CWIC on correction of their own writing, we intended to facilitate as much as possible their noticing the gap between their interlanguage and native speakers’ production, in view of Ellis’s assumption that only the characteristics of linguistic input that are noticed have the potential to be acquired (1995:89). At the very least, we hoped that mastering corpus use would increase the students’ confidence to self-correct and therefore their preparedness to act on their own doubts as they arose. We also expected that a by-product of the kind of apprenticeship we conducted would be to bring certain commonly occurring types of problems to their attention, which they would then look out for.

The broader underlying aim of the text-correcting part of the apprenticeship was to promote a more sophisticated understanding of the role of grammar in conveying meaning and stimulate the students’ curiosity and ability to discover patterns for themselves. The grammatical consciousness-raising benefits of work with concordances have been espoused particularly by Johns (1991b:30). He described the learner armed with a corpus as a linguistic researcher who formulates and tests hypotheses on the basis of evidence from real language use. Thanks to the direct access to data (that is, language examples) that a corpus allows – without the mediation of a teacher, textbook or grammar book – the learner-researcher can engage in an individual process of discovery of grammar and meaning. In this light, we saw the point of grammatical exploration work with a corpus as the development of what Larsen-Freeman called the fifth macro-skill, “grammaring” (1998), rather than the acquisition of knowledge about rules. We also expected that taking on the role of linguistic researcher would help our students go beyond what is dealt with in class, by putting them in a position to question the authority of teachers, textbooks and grammar books and so preparing them for the separation from the security of learned rules that using a language in real life entails.

We began the apprenticeship with a few activities chosen by us to illustrate some of the types of problems that can be investigated. These were conducted by the whole class, working in pairs, and directed step-by-step by the teacher, who introduced applicable software features (such as wild cards and sorting of examples) and often asked leading questions or called attention to particular examples.

Some activities were presented in what Murison-Bowie (1993) called “top-down” form, in that they entailed putting a given rule or hypothesis to the test. For example, the students explored the issue of when to use the definite article with possessive adjectives, after first summarising what they considered to be the rules. From the concordances examined they were able to add extra details to their existing knowledge: they noted *suo fratello* (her brother) but *il suo fratellino* (her little brother); and *mio figlio* (my son) but *il mio amatissimo figlio* (my beloved son). They also detected some misconceptions: both *mia mamma* and *la mia mamma* were found. In the process, some happened to notice examples of *la mia casa* and *casa mia* and reflected on the difference in meaning associated with the different word orders (roughly *my house* and *my place*). In another activity, on near synonyms, the students were first invited to formulate a hypothesis regarding the differences between the verbs *mostrare* and *dimostrare* (which relate to *show* and *demonstrate* but do not correspond exactly) and then to analyse the examples of each in order to refine their hypothesis, with regard to the specific contexts of use. Other activities were conducted in “bottom-up” style; that is, concordances were examined without prior reference to a hypothesis. For example, the students were
invited to observe and classify some concordances of the verb *penso*. After identifying the constructions *penso a* <noun>, *penso di* <infinitive>, *penso a* <infinitive> and *penso che* <clause>, they analysed the meanings associated with each (which included *I'm thinking of* <someone>, *I’m thinking about* <something>, *I expect/intend to*, *I’m taking care of, I’m considering, I think that*) and generalised to make some rules.

After this brief introduction, the students began to use the corpus individually, while revising their own compositions. In order to encourage this, we periodically presented the class with anonymous sample sentences from the previous week’s writing and worked together through ways of using the corpus to make corrections. In this way the students practised formulating questions for investigation, such as “Should I use *infine* or *finalmente* here?” (roughly *lastly* or *at last!*), and devising appropriate searches. When marking their work, we pointed out where they might be able to make corrections with reference to the corpus, and then dedicated some class time each week to individual problem-solving work, with the teacher circulating to give assistance on request.

### 3.2 Text-enriching mode

With the students we referred to this as ‘treasure-hunting’, as it involves exploring the corpus in search of words, expressions and even sentences that can be ‘plundered’ for use in their own compositions. Only a few suggestions in this vein were to be found in the literature on corpus use in language classrooms. For example, Tribble and Jones (1997:80) proposed an activity of fishing for metaphorical language, by searching on the names of colours or parts of the body to see what idiomatic expressions turned up. Then there were activities aimed at finding appropriate collocates, such as Gavioli’s example (1996:142-43) of identifying the adjectives used for winds in Italian weather reports. We saw our treasure-hunting as different from such activities in that it was always conducted with reference to a specific writing task at hand, rather than just to see what came up, and because we presented it explicitly as using *CWIC* to aid the imagination.

A corpus can provide ideas on what to say on a given topic or in a given situation (“What might I say about my holidays?”, “What goes in a job application?”) or how to vary or embellish the language used.

We introduced text-enriching activities quite late in the students’ apprenticeship, and it was in this work that we exploited the corpus as a database of whole texts and a source of frequency lists, as well as the raw material for concordances. In fact, we introduced the students to three types of techniques, one for each of these ‘dimensions’. The first was that of browsing through several texts of a given type. For example, skimming a few film reviews can be useful for drawing up a list of the sorts of points that tend to be made about a film and some appropriate vocabulary for discussing plot, characters, themes or actors. Browsing can also be aimed at finding ways of expressing specific functions. For example, checking just the opening and closing lines of some personal letters or emails reveals various models for apologising or thanking someone. When exploring the opening sequences of just a few personal letters our students enjoyed discovering various non-standard greetings including *Ciao dolcissima Elena* and the thanks conveyed in *Non sai quanto piacere mi ha fatto ricevere tue notizie* (You don’t know how pleased I was to have news of you). These discoveries not only provided them with appealing alternatives to expressions they overuse, but stimulated curiosity and triggered further investigations.
The second technique involved searching on likely key words for the topic or function in question. For example, students writing about moving house might pose questions like “How do you say buy a house and rent a house?”, “What can I put in a description of the house?” or “What can I say about furniture?”. They might begin by searching on casa and appartamento, familiar words that most obviously come to mind. Examining the examples produced and sometimes browsing a few sentences either side would turn up various useful phrases and also further words to search on, such as trasloco (a move) and forms of trasferirsi (to move house).

The third technique was that of perusing frequency lists, drawn either from the whole corpus or a subcorpus of a particular text type. For example, the list of frequently occurring 3-word or 4-word sequences in the film reviews produced many useful combinations. Some of these were terms specific to the field, such as macchina da presa (camera); others suggested observations commonly made about films or directors, like un tema ricorrente and già autore di (a recurring theme, also the author of); and many provided models for the construction of phrases, like un film su and diretto e interpretato da (a film about, directed and played by).

4 Evaluation of this approach and implications

The students’ responses to questionnaires showed they found treasure-hunting sessions particularly rewarding. It seemed they felt more confident about the productivity of work in text-enriching mode, due to its open-ended nature, than in text-correcting mode, where the target was ‘the right answer’. Although they appreciated the value of CWIC for solving grammatical problems, they reported feeling frustrated at times with text-correcting work and occasionally overwhelmed by the arrays of examples.

Since the text-correcting work was evidently more problematic, and had constituted the focus of our training, evaluating students’ performance in that area was our first priority. Given our intention that the training should prepare students for independent use of CWIC outside the classroom, our first evaluation exercise was aimed at understanding the mechanics of their grammatical investigations when conducted without a teacher’s assistance. We videorecorded pairs of volunteers at work with CWIC on set tasks and interviewed them immediately afterwards. Our analysis of their problem-solving processes, based on the transcripts of their conversations at the computer and the interviews, is presented in detail in a separate paper (Kennedy & Miceli 2001; see http://llt.msu.edu/vol5num3/kennedy/default.html).

While many of the students’ investigations were successful, we were interested in what went wrong in those that were not. As was to be expected given the students’ intermediate level of proficiency, a few investigations were derailed by miscomprehension of the examples presented. Some others failed to produce a satisfactory result because the question posed for investigation was not an appropriate or useful one to ask. This problem also often depended on students’ knowledge of Italian, in that the framing of the question reflected a lack of awareness of issues related to the phenomenon under investigation and/or assumptions associated with their first language.
However, in the majority of cases, we identified problems that were not primarily a function of language proficiency. We directed our attention to these because, not only did they have a considerable bearing on outcomes – much greater than we had envisaged – but we saw them as potentially curable through appropriate training. They were of two types.

The first type of problem lay in the students’ behaviour in the role of researcher. Their reasoning processes were often less than rigorous, and some even seemed not to fully appreciate the necessity of working in a logical manner. This was accompanied by an evident ignorance of possible pitfalls. There were many different manifestations of these difficulties. For example, in some cases students did not take care to tailor a search strategy to the exact question being dealt with. Some showed a propensity to base the conclusions of their investigations on assumptions rather than the evidence collected from the concordance examples. In cases where no examples of a particular form were found the conclusions drawn were often quite unrealistic, such as “So the answer is not $a$. Therefore it must be $b$”, when no evidence concerning the form $b$ had been found or even sought. Prior assumptions also sometimes overrode or corrupted a rule they had just discovered, when they came to apply it in their own sentences. At times, too, students neglected to ensure they took into consideration only the examples that were closely analogous to the case in question. For example, in deciding whether an article is required before *lunedì scorso* (last Monday) several students took as models the examples containing *il mese scorso* or *la settimana scorsa* (last month, last week), both with article, overlooking those containing *<day of the week>* scorso with no article. There were also a few cases where students simply chose to ignore examples that did not match their expectations or support their hypothesis.

It seemed we had underestimated the need to train the students as researchers, due to our own unstated assumption that the necessary attitudes and reasoning habits would be intuitive. We therefore decided to change our approach to training the students in text-correcting work, by shifting the focus from simply illustrating the types of lexical and grammatical questions that can be addressed to developing awareness of what the researcher’s role entails and the principles and processes involved in formulating and testing a hypothesis. Central to this would be the practice of engaging the students in class in a systematic reflection on the steps involved in an investigation, typically:

1. formulate the question
2. devise a search strategy
3. observe the data and select relevant examples
4. draw conclusions.

In order to draw their attention to the likely pitfalls in each step, we decided to also incorporate exercises in the style of ‘spot what goes wrong’ and collectively draw up sets of tips for avoiding them.

The second type of problem had to do with the students’ effectiveness in observation of the data, an essential component to any research exercise. The failure to notice patterns in concordance examples, and their similarities and differences, was a prime cause of unsuccessful or only partially successful investigations. The students often overlooked word patterns or concentrated their attention on only part of a pattern, not perceiving related words to the right or left and, more importantly, not seeking them out. While they recognised common idiomatic expressions as clusters of words that belonged together, and certain combinations of verb and preposition, they seemed to lack...
sufficient awareness of the myriad other bonds linking words into patterns like lunedì scorso (last Monday), con grande sorpresa di (to [someone’s] great surprise), tutto ciò che (everything that), niente da fare (nothing to do) or per motivi di lavoro (for work reasons). Furthermore, most lacked curiosity about any patterns other than those they were expecting to find, and therefore did not further explore and exploit them.

So our second concern in redefining our approach to training was to raise the students’ consciousness regarding observation. We sought ways of encouraging them to look at examples with a more open mind towards the possible bonds between words, and of making them more alert to differences between what they assumed they would see – often the effect of first language interference – and what was actually present in the examples. Reflecting on the problem of noticing words and their relationships led us to explore Nattinger’s discussion of “lexical phrases” (1980) and Lewis’s “Lexical Approach” (1993, 1996), which stressed the importance of multi-word items as building blocks of language, alongside single words. Lewis’s claim that “much of our supposedly ‘original’ language use is, in fact, made of prefabricated chunks…” drawn from a kind of personal mental database of formulaic language (1996:10) triggered a reassessment of the significance of treasure-hunting in our apprenticeship. We came to think that it might be much more valuable to learners than we had imagined, by virtue of encouraging them to identify and consciously borrow chunks.

We therefore decided to rearrange the apprenticeship by introducing treasure-hunting first and giving it much greater emphasis. In this way we mean to highlight, from the outset, the function of the corpus as a database of models of whole texts, sentences and prefabricated chunks. We hope to encourage the 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. They should then be better armed with observation skills for attacking grammatical problem-solving, shifted to the second stage of the apprenticeship.

We have also abandoned the net distinction between text-enriching and text-correcting modes of using the corpus, which saw treasure-hunting techniques employed only for text-enriching purposes. We now consider that a treasure-hunting mentality, associated with the practice of borrowing a given model, should be beneficial in text-correcting work too. That is, it may be valuable to adopt an open-ended approach in dealing with some grammatical problems, especially for our intermediate learners who, as noted above, sometimes have difficulty setting up an investigation because they do not fully appreciate the terms of the problem. For example, one student wanted to say to the man’s great surprise and posed herself the question of which preposition to use for to in that situation. She looked up sorpresa and found examples including con grande sorpresa dei professori and con nostra grande sorpresa. She therefore inserted con into a phrase of her own construction, which included an unwanted definite article, producing con la grande sorpresa del signore. If she had adopted a treasure-hunting approach instead and asked herself “what is an expression for to someone’s great surprise?”, she might have been more likely to borrow and adapt the whole phrase, to produce the correct version con grande sorpresa del signore.
5 Conclusion

A key change in the way we conduct the students’ apprenticeship with CWIC is to introduce them to the corpus with an ‘observe and borrow’ mentality first, before progressing to an ‘observe and derive rules’ approach such as we had originally viewed as the norm for corpus investigation. The next step is to evaluate this new style of apprenticeship, with particular attention to the effectiveness of the students’ observation and borrowing. From their comments on class work, we already have some evidence of students employing a treasure-hunting mentality in their problem-solving and, as a result, discovering more than they set out to look for or reformulating the question to be dealt with. For example, one student described an investigation as follows: “I wanted to make sure I was using this word [invece] correctly and in looking it up found an expression that was much better than the one I had [in mind]”. As in the first evaluation exercise, we will focus on understanding the mechanics of their work with the corpus and identifying ways to improve the training.

Since we embarked on this project, production of larger corpora of contemporary written Italian has been under way in Italy. While we expect these to be extremely useful, and will also be experimenting with them when they come on-line, we believe our experience has confirmed the value, for intermediate learners, of working with a relatively small corpus containing texts of familiar types that can provide useful models for their own writing.

Acknowledgements

We are indebted to Catherine Kennedy for valuable feedback on an earlier draft of this paper.

CWIC was developed with the support of the Australian Government’s Committee for University Teaching and Staff Development (CUTSD) and Griffith University.

References


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 The email lists and magazine columns proved to be particularly valuable sources. 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, 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.

3 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.

4 The paper cited was based on a presentation at the fourth conference on Teaching and language corpora, held in Graz, 19-24 July 2000. See http://www-gewi.kfunigraz.ac.at/talc2000/Htm/index1.htm