Cultivating effective corpus use by language learners

Claire Kennedy and Tiziana Miceli, Griffith University

Claire Kennedy (corresponding author)
Cassamarca Senior Lecturer in Italian Studies
School of Humanities, Languages and Social Science
Griffith University
Nathan QLD 4111 Australia
c.kennedy@griffith.edu.au +61-7-3735 7141

Tiziana Miceli
Lecturer in Italian Studies
School of Humanities, Languages and Social Science
Griffith University
Nathan QLD 4111 Australia
t.miceli@griffith.edu.au +61-7-37357075

Abstract

While there is widespread agreement on the expected benefits of hands-on access to corpora for language learners, reports abound of the difficulties involved in realising those benefits in practice. A particular focus of discussion is the challenge of transferring the skills of the corpus linguist to learners, so that they can explore this type of monolingual and unmediated resource effectively in a ‘learner-as-researcher’ or ‘learner-as-detective’ role. In this paper we present a positive experience of an apprenticeship in corpus use, which we attribute to a training approach that seeks to exploit the unmediated nature of corpora rather than treating it as a problem for learners. Our approach is not aimed at equipping learners with skills that will allow them to derive rules or explanations from the systematic analysis of corpus data. Instead we aim to inculcate in our students a propensity for open-ended searches, and an observe-and-borrow-chunks mentality, oriented to copying models from a corpus to enrich and enhance their written production in particular. We see the training process not only as supporting their development as effective corpus users but also, in encouraging their curiosity about the language patterns to be found and their authentication of the use they make of those patterns, as likely to be beneficial to their ongoing development as strategic independent language learners more generally.

Key words: corpus, foreign-language learning, independent learning, reference resource, writing, life writing

Introduction

... it's made me feel like I can do anything. ... Yea, it's made me feel like that I can write to those [people] in Italy and not be afraid that I've said something really stupid. ... It's definitely made me love to learn Italian more ... It's opened up the door to so many um, like, I can do so much more learning on my own now. Before I was relying too much on the teacher and the grammar exercises we were given and now I rely more on these [online dictionaries] and [the corpus], to check what I'm doing. So it means that I could continue with Italian after I finish [my degree] more than if I didn't have that resource (Student 'S5', interview)

This kind of reaction to our introduction of a new language-learning resource epitomises what we are working for as teachers: the student finding that she loves learning (even) more, and feeling empowered and confident to go forward as an independent learner. The positive experience of this student and several of her classmates encourages us and reinforces our view that use of a monolingual corpus, such as our *Contemporary Written Italian Corpus* (CWIC), can be highly productive for language learners, even when their proficiency is only at intermediate level. Yet a recurring theme in the literature in this area is that, while there is broad agreement on the potential benefits of corpus use for language learners, there has been remarkably little take-up, due to various types of challenges that arise; indeed, reports of successful ‘direct’ - or hands-on - applications are rare. Our aim in this paper is therefore to argue for the benefits of the specific training approach we have developed, in overcoming or obviating such challenges and promoting the kinds of rewards that the student quoted above perceived. We discuss an investigation into our students’ use of CWIC, in which we sought a nuanced understanding of what they could achieve with it, what made some of them particularly effective users, and how the experience supported their development as language learners generally.

Rationale for our approach

For several years the application of corpora in language courses has been no longer the prerogative of teachers with a technological bent, who prepare and guide activities for the learners; instead each learner can have direct and rapid access. The technological transformation of everyday life and communication has meant there is nothing exotic about online searching, and learners now come equipped with digital skills and intuitions about querying that can be tapped in corpus interrogation (Pérez-Paredes et al., 2011). Various monolingual and bilingual corpora are freely available and exploited in diverse learning contexts.

Investigations of L2 learners’ work with direct access to corpora have identified a range of positive reactions, which can be grouped in two main categories: appreciation of the type of learning promoted and the type of language encountered (Chambers, 2007). There is abundant evidence of students engaged in self-directed corpus activity feeling more responsible for their learning and highly motivated (Chambers, 2005; Geluso & Yamaguchi, 2014; O’Sullivan, 2007; Yoon, 2011; Yoon & Jo, 2014). As for the type of language, students have reported that they value dealing with real language as opposed to invented examples in a textbook (Chambers, 2005; Yoon & Hirvela, 2004; Yoon, 2011) and, in particular, engaging with ‘chunks’ of language in use (Geluso & Yamaguchi, 2014; Yoon & Hirvela, 2004).

Similarly, many teachers’ enthusiasm for corpus applications is linked to research on phraseology and the role that lexis plays in language learning. A rationale for promoting learners’ use of corpora can draw on Sinclair’s (1991) idiom principle and Hoey’s (2005) theory of lexical priming (Poole, 2011), and borrow practically from Lewis’s (1993) ‘Lexical Approach’ with its principles and practical guidelines for teachers in developing learners’ perception and use of word chunks as building blocks of language. For the application of corpora to language learning challenges any traditional view of vocabulary as a collection of single words with fixed meanings; it offers learners the opportunity to critically observe words in use, in relation to their cotext and context (Geluso & Yamaguchi, 2014; Sha, 2010).

However, further empirical studies are needed on what students actually achieve with corpora, to support the evidence of students’ favourable perceptions (Yoon, 2011), as some

researchers have reported strongly negative learner reactions (Chambers, 2007; Chang, 2014). Indeed, several scholars have argued that use of corpora in the language classroom remains limited, and less successful than expected (Ädel, 2010; Breyer, 2011; Chambers, Farr & O’Riordan, 2011; Frankenberg-Garcia, 2012; Kılıçkaya, 2015; Poole, 2011; Römer, 2008, 2010; Tono, Satake & Miura, 2014). In this vein, Breyer (2011, pp. 3-4) draws attention to the persistence of a “gap between research efforts and actual application in teaching practice”. Römer (2010, p. 18) feels “hesitant” to say that corpora “have after all fully ‘arrived’ on the pedagogical landscape” given that relatively few teachers and learners seem aware of their availability, while Tono et al. (2014) remark upon many teachers’ lack of experience and training in using corpora. Tribble (2015, p. 57), reporting on his recent survey of language educators, confirms that “the reasons for not using corpora emphasise the still widespread lack of knowledge about corpus tools”.

In the discussion of why corpora are not more readily taken up by language teachers and learners, three key problems have been identified, each of which we have sought to address in our own approach to supporting ‘hands-on’ corpus use by our students. First, several scholars (Ädel, 2010; Braun, 2007; Breyer, 2011; Chang, 2014) have pointed to a lack of pedagogically oriented corpora and software. Initially, most available corpora had been created for linguistic research purposes, with content unlikely to suit learners’ needs (Braun 2007), while the searching software tended to be unnecessarily complex, and searches often generated irrelevant data because of the size and content of the corpora (Breyer, 2011). The situation has improved in recent years thanks to tools such as *Sketch Engine* which facilitate creation and management of corpora by teachers and learners themselves (Kilgarriff et al. 2004). We use a 500,000-word corpus called CWIC (*Contemporary Written Italian Corpus*), compiled by us specifically for our students and their university learning context, with a custom-built search engine and user interface. We have integrated training in use of CWIC into the existing creative-writing workshop of a second-year, second-semester - i.e. intermediate-level - course. Over the semester, each student writes an autobiography in five chapters and posts it on a personal wiki within the course site so as to share the process and product of her¹ writing. This writing project (for assessment) is structured in 5 *compiti* or homework items, each of which includes not only the writing of a chapter but also reflective activities, such as self-correcting the text, reporting why and how the student used the corpus, and providing peer feedback to other students. We begin the corpus ‘apprenticeship’ four weeks into the course, before the students start on *compito* no.3, introducing CWIC as a reference resource to support the autobiographical writing task, available to be consulted autonomously in search of language patterns useful in articulating their ideas and feelings. CWIC consists of short written texts on everyday topics by native speakers, largely from emails and magazines (see Kennedy & Miceli, 2002 for a description of the text types and sources, and the selection process).

The second area of discussion around limited corpus use in the classroom concerns the difficulties that learners have in behaving like linguistic researchers when using corpus data, and the difficulties that teachers have in training them in the necessary skills (Ädel, 2010; Braun, 2007; Breyer, 2011). However, as a result of our experience over time in training students, we now see the assumption underlying this debate - i.e. that classroom concordancing requires transferring corpus-linguistic research skills to language learners - as problematic. Behaving like a researcher requires more than stringent reasoning; ideally, like a

¹ For brevity, we use ‘she’ and ‘her’ as generic pronouns for individual students.

chemist interpreting data obtained from laboratory measurements, a researcher seeking to derive information or rules from corpus data should apply the following:

- a) rigour in setting up the experiment, to ensure no relevant data will be ignored
- b) rigour in observation and interpretation of the data produced, which entails:
 - accounting for all relevant data
 - interpreting, labelling and classifying each data item accurately and completely
 - recognising similarities and differences between data items
 - identifying and catering for all factors that may affect the manifestation or otherwise of data items, and their interpretation.

Success therefore requires familiarity with the type of data, knowing what to look for and be wary of; that is - in the case of corpus research - *high proficiency*.

The basis of our approach is therefore to *downplay* the learner-as-researcher notion, and instead seek to cultivate in learners a propensity for open-ended searches, and an 'observe and borrow chunks' mentality. Our approach entails presenting a monolingual corpus such as CWIC as an aid to the imagination and memory, and is aimed at enabling learners to make effective use of it to enhance their written production without requiring the rigour and linguistic proficiency of a corpus-linguistic researcher. In line with this, our focus, in introducing CWIC, is on extending the students' repertoire of reference-resource functions, to add those that we see a monolingual corpus as lending itself to in particular. We do this by presenting two key ways of using CWIC: to enrich the content and language of a text through what we call *pattern-hunting*, and to edit a text for lexico-grammatical accuracy through *pattern-refining*.

By *pattern-hunting* we mean exploring the corpus for ideas and language patterns - i.e. groups 'of words - to borrow, adapting them as necessary. L2 writers can find themselves at a loss for ideas on what to say on a topic, or lacking in words to express their ideas, because they cannot imagine certain expressions that they have not encountered or noticed, or because ways of saying things in their L1 do not have a predictable equivalent in the L2. So, we start our students' CWIC apprenticeship with activities that entail hunting for answers to open-ended questions like 'What might I say about my university life?', 'How might I describe my love relationship?' and 'How might I sign off an email affectionately?'. We explore three techniques for seeking word patterns that our search engine supports: a) browsing through whole texts, chosen on the basis of text type and optionally title; b) searching on words likely to be associated with the topic concerned and perusing the concordance lines displayed (and optionally their whole source texts); and c) scanning frequency lists for common 2/3/4/5-word combinations. The very first class activity involves hunting for useful patterns for writing about love and relationships.

Later in the apprenticeship, we introduce students to ways of editing their text for lexico-grammatical accuracy through *pattern-refining*. In contrast to pattern-hunting, which involves open-ended questions, pattern-refining work is specific problem-solving with the corpus; it is aimed at finding models for patterns when you *do* know what you want to say and know one or more component words of the target pattern. The technique used is almost always that of searching on a word or combination of words.

By concentrating on *pattern-hunting* and *pattern-refining* in the apprenticeship, we aim to raise our students' awareness that much of language use is in prefabricated chunks drawn from a kind of mental database of formulaic language (Lewis, 1996), and instil a view of CWIC as a source of models for such chunks, along with a habit of borrowing rather than

inventing them. We hold that these functions can be productive and rewarding for intermediate-level learners and can acquaint them with corpora as reference resources in a way that stands them in good stead for *later* developing into fully-fledged ‘language-learners-as-researchers’.²

As a result of our past evaluations, we now include in the apprenticeship an in-class reflection forum towards the end of semester, in which students share and compare their ways of using CWIC and other reference resources (online dictionaries, translators, conjugators, discussion forums, grammar-explanation sites, search-engines, plus grammar books - the only paper-based resource usually mentioned). Through this explicit discussion we also seek to raise their awareness that effective CWIC use requires understanding that:

- each type of reference resource (e.g. monolingual dictionary, bilingual dictionary, monolingual corpus, verb conjugation manual) lends itself to distinct functions
- pattern-hunting and pattern-refining are functions that exploit the unmediated and monolingual nature of the CWIC corpus
- the characteristics of a corpus may limit its usefulness for certain functions (e.g. too small, too narrowly focused) but not necessarily others
- working with corpora requires preparedness to proceed by trial and error and acceptance of the uncertainty of finding a satisfactory answer
- the development of skills in corpus use is a long-term process which goes well beyond a one-semester apprenticeship.

A third issue that has been raised in relation to the application of corpora in the language classroom is the need for students to authenticate corpus language (Braun, 2005) and for teachers to create a learning environment that enables students to do this (Breyer, 2009). Since texts lose their original context - a core feature of their authenticity - once inserted into an electronic corpus³, corpora pedagogy needs to “create conditions that enable learners to authenticate corpus data through some sort of engagement with it” (Mishan, 2004, p. 222). Our response to this lies in the decision to introduce the corpus as a resource that supports a writing task that engages the students personally, presenting its use as an integral part of that writing process. Students obtain from the corpus word patterns that are valuable to them individually, for their writing about their own life experience, and from *compito* 3 onwards they are invited to accompany each autobiography chapter with an account of their CWIC use. In this way, they are encouraged to engage in “a process of authentication” (van Lier, 1996) of the *use* they make of the language they find in CWIC’s texts. Furthermore, we see the in-class reflection forum as not only about sharing ideas but helping students develop, and take pride in, their own individual ways of using the corpus - a process of ‘making it their own’. We had identified this as a key discriminating characteristic between effective users and others in a previous study (Kennedy & Miceli, 2010).

Aims and method of current study

Aims

² Geluso and Yamaguchi (2014) also introduced their students to corpus work through pattern-hunting, and reported positive outcomes in terms of the students being able to use the ‘formulaic sequences’ they found, in conversation.

³ The problem of decontextualisation is to some degree ameliorated in our case by the fact that our search engine displays the source text identification (and type) for each concordance line, and users can click on any concordance line to immediately view its complete source text.

Our increasing confidence, in recent years, that substantial numbers of students in each annual cohort seem to make effective use of CWIC, led us to conduct an investigation among one cohort, aimed at understanding in some depth what they were able to achieve with it. We posed two initial questions:

- Q1 What do our students use CWIC for and with what success in enhancing their written compositions?
- Q2 Beyond the practical outcomes, what characterises the work of those we see as effective corpus users?

Our analysis of data in addressing these two questions led us to reflect on two further questions:

- Q3 What are the broader benefits of this experience with a corpus for the students' development as language learners?
- Q4 In what ways does our specific approach to training support these learners' development as effective users of a monolingual corpus?

Our interest here is in those students who have embraced work with CWIC. The question of why others have not - the minority in the cohort discussed here - will require a separate study.

Participants and method

The 24 students in this cohort who passed the course - 19 women and 5 men - are referred to in this paper as S1 to S24. Prior to enrolling in this course, 14 had taken three semesters of Italian with us, having joined our program as complete beginners, and 10 had entered the program more recently, with prior learning at school or elsewhere. While 10 were enrolled in a Bachelor of Languages and Linguistics, 4 were in Humanities degrees, 3 in Education, 2 in Business, and 1 each in Music, Fine Arts, International Relations, Political Science, and Journalism.

Data were collected from three sources: all students' accounts of CWIC use accompanying their autobiography chapters (i.e. *compiti*) 3, 4 and 5, submitted for assessment in weeks 7, 10 and 12 of the 13-week semester; the reflection forum in week 10; and individual interviews with selected students after the end of the course. The first two of these sources were integral to the apprenticeship, as outlined above, while the interviews were conducted solely for research purposes. The forum was held in class time, in two separate groups of similar size. Each group's session lasted one hour and was co-conducted by both authors. The sessions were semi-structured, as we followed a detailed plan of activities and open questions about the students' use of CWIC and other reference resources, but several points not anticipated in our plan were raised by the students and discussed.

We invited for interviews the students we considered to have particularly embraced *CWIC* use, as the purpose was to capture an in-depth picture, following up on the students' accounts of CWIC use accompanying their *compiti* and their observations in the forum. Of these, five (S1, S2, S3 S4, S5) were available for an interview within our chosen time frame, four women and one man. Most were high achievers in the course: they attained grades of 7, 6, 6,

4, 7 on a 7-point scale, respectively.⁴ Four intended completing an Italian major, the fifth was finishing Italian studies with this course.

The interviews were conducted just after the publication of results at the end of the course, while the students' memories of the work completed were fresh but the risk of their perceiving encouragement to answer in ways that pleased us was reduced. Each interview was conducted face-to-face between the student and one or both authors, lasting approximately one hour. They were semi-structured - based on the set of open questions in Figure 1, but with each taking unplanned turns in response to the student's input. We opened with a general question asking the student how she would describe CWIC to a newcomer, in order to initiate a conversation driven as much as possible by the student, through which her views of CWIC and other reference resources could emerge gradually. We left certain questions to the end, as they could have conveyed expectations or judgments on our part, and therefore have influenced responses to any subsequent questions.

[Figure 1 goes here]

The forum sessions and interviews were audio-recorded and the recordings transcribed by an assistant. The transcripts were checked and corrected by one author.

The results we present below in relation to Q1 (our description and classification of students' CWIC operations) are based on our interpretation of the students' retrospective self-reporting, through the brief written accounts accompanying their *compiti*, and their oral accounts in the forum and interviews. We cannot claim to provide either an exhaustive report or a quantitatively representative sample of the students' work with CWIC during the course, but are limited to illustrating some key ways the students *appeared* to us to use it. There were no feasible mechanisms for capturing completely the use the students made of CWIC (including searches they deemed not useful) at any time and from any place, let alone their intentions and thought processes. Self-reporting operations in such detail would have been too onerous, and we had no option for collecting such information automatically from user logs.

Of the three main techniques we had introduced, most students only reported using word or word-combination searches, hereafter referred to as 'lookup' operations, and we focused our analysis on those.⁵ Figure 2 summarises our approach to classifying them.

We could not classify each operation as 'pattern-hunting' or 'pattern-refining' *per se*, because of the ways they sometimes developed: for example, an operation might begin with a pattern-refining intention but produce a pattern-hunting type of outcome, with a pattern borrowed that was not predictable at the start. Instead, we classified the *initial* aim of each operation as open-ended (pattern-hunting) or specific (pattern-refining). We then assessed each operation as to whether it was valid and whether any patterns were successfully borrowed as a result.⁶

⁴ Passing grades awarded for courses at our university range from 7 (High Distinction), to 4 (Pass).

⁵ Only one student also reported scanning frequency lists of 3-4 word patterns (S23), and two reported browsing whole texts (S5, S7), in their *compito* accounts, although two others mentioned, when interviewed, that such browsing had sometimes been useful (S1, S2).

⁶ The types of errors we treated as candidates for labelling 'surface' errors were: absence of definite article where required or presence where not required; lack of grammatical agreement for gender, number or person; slightly incorrect verb form; mis-spelling. Actually labelling an error as 'surface' then depended on the aim of the operation concerned; for example, the absence of a definite article was not merely a surface error if the aim of the operation concerned was to determine whether or not an article was required in a certain pattern. In most

[Figure 2 goes here]

Our analysis in relation to Q2 (on the characteristics of effective users) is the fruit of our reflection on the students' lookup operations and their comments in the five interviews and two forum sessions, primarily, although we also checked the *compiti* for any applicable comments. Each author independently examined the transcripts, flagging utterances that we considered identified attitudes and behaviours that had helped render the students' CWIC use effective or ineffective, and/or suggested practical steps we could take to improve our training. We then compared our results and reconciled through discussion the few differences of interpretation that emerged. Given the manageably small amount of textual data - transcripts of approximately seven hours of talk - it was not necessary to use any software to organise the data and facilitate this analysis.

Results and discussion

Q1 What do our students use CWIC for and with what success in enhancing their written compositions?

The *compiti* provided evidence that over two thirds of the students - 17 out of 24 - made productive use of CWIC for lookup operations, in that they conducted valid operations and borrowed some patterns in their compositions. Twelve students conducted some operations with open-ended initial aims (S1, S2, S3, S4, S5, S7, S9, S16, S17, S19, S20, S23), three of whom also used operations with specific initial aims (S1, S4, S5), while five used operations with specific initial aims only (S13, S15, S18, S21, S22). Of the remaining students, one provided reports of CWIC use that we disregarded as incoherent while the other six provided no account of CWIC use.

The five interviewees were among those we considered the most productive CWIC users, in that CWIC had evidently been of significant help in enhancing their compositions. Three of them were particularly productive in operations with open-ended aims (S2, S3, S4) and the other two in operations with specific aims (S1, S5). It became clear in the interviews that four of them (S1, S2, S3, S5) had used CWIC much more than they had reported (especially after *compito* 3, when reporting was no longer explicitly requested) and indeed by then used it habitually.

The examples in Table 1 illustrate the kinds of successful outcomes achieved by the students in using lookup operations in CWIC to enrich and render more appropriate the language in their compositions. The initial aims of these operations ranged from very open-ended to very specific; that is, from

- looking for something to say about a topic (nos 1-3)
- through
- looking for words/patterns associated with a particular word, in the broadest sense of checking which potentially useful patterns occur nearby in the same text (nos 4-6)

cases the surface errors resulted from neglecting to adjust all necessary noun and adjective endings for grammatical agreement.

and

- looking for ways to describe something or say something specific about something (nos 7-9)

to

- looking for examples of a word/pattern used in a sentence (nos 10-11)

and

- looking for specific expressions including the searched word(s) in a polyword, collocate or fixed or semi-fixed expression (nos 12-14)

and very specifically

- looking for evidence confirming that the searched word/pattern exists or can be used with a certain meaning in a certain context (nos 16-17).

In operations with specific initial aims, the outcome(s) sometimes diverged from the aims, in that the student noticed and borrowed something instead of, or in addition to, what she was initially seeking (e.g. 10, 13, 18).⁷ In these cases, and in operations with open-ended aims, the patterns borrowed often did not include the searched word(s) but were found in the cotext in a concordance line or by perusing the whole text that a concordance line came from (e.g. 2, 3, 7, 19).

While the patterns borrowed were often polywords, collocates, or fixed or semi-fixed expressions (e.g. 6, 10, 12, 13) - with reference to Lewis's (1996) lexical chunk types - they were sometimes not classifiable as any of these but just unique combinations of words that appealed to the student for her purposes at the time (e.g. 3, 4, 5, 8). The borrowings varied in length (from 1 word to 7, with many of 4-5 words) and in the types of changes needed for integration into the composition (from no change, through straightforward changes of gender, person or tense, to significant changes such as separating pronouns from verbs and changing word order) and in the degree of accuracy achieved in this adaptation.

The students described some of these operations as very quick, when they spotted something they wanted in the first couple of concordances looked at. Others involved scanning or carefully sifting through several concordances, or flicking through a few screenfuls and scanning only the first couple in each, in search of inspiration.

These examples highlight the rewards that some students obtained from posing open-ended questions for their lookup operations and/or being open-minded in observing the data produced by a search, in the sense of being prepared to borrow beyond their original intentions by 'mining' the cotext of the searched word(s). An 'open' approach often allowed the outcomes to exceed their initial expectations.

In our view, the students who demonstrated neither type of openness missed valuable opportunities.⁸ This is illustrated by S22, who described her aims generally as "confirm[ing] the use of words I found and ... see[ing] if it was used in a similar way to the way I wanted to use [it]" and, in one case, "... to check that [two words] were used interchangeably". Describing the room she had lived in throughout her childhood could have provided ample scope for open-ended searches and rich borrowings, yet her searches were limited to the kind

⁷ Clearly, in cases of open-ended aims with a borrowing outcome, it makes no sense to consider whether the borrowing outcome matches the aims.

⁸ In commenting on the students' work in this way, we are assuming that they did not do certain searches and/or did not notice certain patterns. It may be, of course, that they *did*, but did not wish to use the patterns noticed and did not report having observed them.

of checking that the adjective ‘accogliente’ (welcoming), found in a dictionary, can be applied to a house.

The examples also illustrate another benefit of a chunk-borrowing orientation to the corpus, which is especially significant for our students who are at an intermediate level of proficiency: unlike a researcher, a chunk-borrower does not need to understand and account for all relevant examples presented by a lookup. The operations in these examples were almost always such that the student could have considerable confidence in basing a borrowing on her analysis of only the concordance lines she found comprehensible, without a high risk of making a misjudgment by ignoring those she did not readily understand. Indeed, in the great majority of the students’ accounts of borrowings, they made reference to only one concordance line on which the decision to borrow was based.

In line with this, we noted that productive work with CWIC was not limited to the high-achieving students, as the 17 productive users had grades distributed across the full range for passing the course, while the students who did not provide comprehensible evidence of CWIC use spanned all passing grades except the highest. Indeed, one interviewee, S2, asserted strongly: “I don’t think you need to be, like, a ‘7’ student to be able to use CWIC. I think it’s actually helpful for people [like me] who struggle a little bit”.

[Table 1 goes here]

Q2 Beyond the practical outcomes, what characterises the work of those we see as effective corpus users?

Through our analysis of the students’ CWIC operations and attitudes, we identified three key characteristics of effective CWIC users. First, effective use reflected conscious use of an ‘observe-and-borrow-chunks’ mentality such as we had sought to instil: an expectation of borrowing pre-fabricated chunks where possible rather than assembling them oneself, coupled with preparedness to notice and borrow patterns beyond those initially sought. We found this mentality observable in the three key stages of the CWIC lookups: posing the question, devising a strategy for addressing it, and examining the data. *Posing the question* in a way that avoided making assumptions about the data to be found was evident, not only in operations with open-ended aims, but also many with specific aims. We saw it as a matter of adopting a ‘lexis-oriented view’, alert to the unpredictability of the length and structure of patterns, which generated questions along the lines of “What words do I use with ... to say ...?” - rather than a ‘grammar-oriented view’ likely to generate questions such as “What preposition do I need with ... to say ...?”. For instance, in example 12, S5 posed the question “How do I say ‘I come from a family of five’?”, rather than “What preposition do I need in ‘a family of five’?” with its attendant assumption (that the problem is simply one of discovering which preposition to plug into the pattern ‘*famiglia* <preposition> <number>’). Then, in *devising a strategy for addressing the question* a chunk-aware mentality was manifest when students searched on only one or a few words with a view to borrowing several, deliberately fishing for a pattern whose form was not predicted (e.g. 13, 14, 18). Similarly, in *examining the data* the students demonstrated a chunk-aware mentality when they looked to both right and left of a searched word in a concordance line, without assuming that it would be bonded only in certain ways or with a certain number of words (e.g. 2, 3), or that the only useful patterns would be those that actually included the searched word(s).

Importantly, there was evidence that the interviewed students were *conscious* of approaching CWIC with an ‘observe-and-borrow-chunks’ mentality. S2, who had a propensity for open-ended aims in her lookups, often seeking ideas for content as much as wording (e.g. “I wanted to say something about my dog”) indicated that she saw having an open mind about what she might find as essential to her approach to CWIC and its usefulness:

because with just the dictionary and the grammar book and stuff ... it’s more structured. You know exactly what kind of thing you’re looking for. But with CWIC you’re looking up a word or you’re looking up something because you have no idea or because you don’t really have that certainty

Yet this was also true of S5, although she always had a specific initial aim for her CWIC lookups, and noted in interview that she did not see the point of doing searches “just to see what came up”, as had been recommended by other students in the forum. A preparedness to notice and borrow patterns that she had not set out to find was clearly part of her habitual way of working: “...there were times when I was looking for something to do with a certain word and I would find something else and I then tried to add that one in”.

Evidence that this mentality was conscious also comes from statements by four of the interviewees (see comments 1-4 in Table 2) conveying awareness that the value of CWIC searches lay essentially in looking for how a native speaker (NS) might express something, understood as likely to be different from how they as non-native speakers (NNS) might assemble the pattern themselves.⁹ And the concept of ‘what a NS might say’ referred to a phrase, a sentence, or ‘something I wanted to say’ that they sought *as a whole* rather than a series of words. Furthermore, S3 explicitly linked this point to her own experience of the difference between using CWIC and Google translate (comment 3). While the lack of 1-1 correspondence between words or patterns in two languages is something we seek to convey to students from the beginning, S1 highlighted that appreciation of this does not necessarily come easily (comment 4). A final item of evidence (see comments 5-7) is three interviewees’ demonstrated awareness that the examples displayed by a search come from whole texts, which are easily accessible by clicking on concordance lines displayed, or via the index of texts by text type, for browsing (or ‘trawling through’, as S1 put it).

[Table 2 goes here]

The second characteristic of effective CWIC users that we extracted from the interview analysis was an appreciation of the type of reference resource it is, and of its functions, merits and limitations in relation to those of other resources, as far as the students’ individual learning purposes were concerned. Each interviewee expressed clearly the functions she used CWIC for, and why (e.g. comment 8). Of particular interest was S1’s description of the arrival of CWIC as “filling a gap” in her resource kit, giving her access to a function she had not realised she was lacking but embraced once discovered: “a good source of reference... for the way a word would be used ... My need to know how it goes in the sentence” (comment 9). Each interviewee also showed awareness that others used CWIC differently, in ways they appreciated as valid even if they did not suit their own learning preferences (e.g. comments 10, 11).

A merit that all the interviewees attributed to CWIC was authoritativeness: that the examples were reliable because they came from NS-written texts which had been vetted by us and whose authorship and origin were known. They noted that examples provided in dictionaries

⁹ The fifth, S4, did not explicitly convey this in the interview and it would have required leading questions to elicit it from her. We perceived her work with ICL to be informed by this idea but we cannot tell to what extent she was aware of it.

and forums were often invented (S3, S5) and lacked contextual information (S1, S4); and the products of Google Translate probably incorrect, or like “a guess” (S1, S3); and material found by Google searching possibly the product of NNSs (S4, S5) or of Google Translate operating on an original site in English (S1, S5) and anyway often from unknown sources (S1, S2, S4, S5).

Some interviewees cited relevance to their needs as a particular merit of CWIC, and this was associated with the custom-built nature of both the corpus and its user interface. S3 and S4 lamented the amount of irrelevant material of different types they had to wade through if using Google to search on a word or pattern of interest (comments 13 and 14). While the small size of CWIC helped make it manageable in this way, the size was also the one limitation of CWIC that was referred to in the interviews. S5 raised this (comment 12), and she saw it as an increasing problem, at least for the pattern-refining function that she mostly used it for.

The interviewees’ comments on search and display parameters in the user interface (see Table 3) were interesting because they showed not only their sophisticated understanding of the functions they were using CWIC for, but also the way each personalised her use of CWIC to suit her individual learning style.

[Table 3 goes here]

The third characteristic of effective users that we identified in the interview analysis was that of being undeterred by the uncertainties of working with CWIC, as a monolingual and unmediated resource; that is, one that provides neither translations nor explanations (Frankenberg Garcia 2005) - and with which it may therefore be difficult or impossible to find a definitive answer to a particular question. When we asked in the interviews “Do you ever get frustrated with CWIC searches or feel bogged down?”, we envisaged two key factors as likely to generate such feelings: being confronted with many examples that are not readily comprehensible; and getting nowhere even after trying various ways of addressing the aim of an operation. The interviewees all reported what we saw as a beneficial attitude - of not succumbing to frustration or feelings of inadequacy when engaged in operations that did not produce results easily, or at all. Comments 15 and 16 in Table 2 by S4 and S5 concern the two anticipated factors specifically. S2 reported a personal technique to avoid feeling overwhelmed by the examples: she set the display parameter to show only 10 concordance lines at a time. She also showed (comment 17) she was comfortable with ignoring examples that did not appeal to her. Interestingly, S1 actually saw herself as less prone to frustration with CWIC than other resources because she could keep changing the way she searched, varying the searched word(s) and the search and display parameters, until she found something that she both understood and deemed useful (comment 18), and because “[it’s] a corpus that’s very easy to search”.

Q3 What are the broader benefits of this corpus experience for the students’ development as language learners?

When considering the ways our students made effective use of CWIC and the interviewees’ attitudes to it, we see positive implications of this apprenticeship with a monolingual corpus for their development as language learners generally. First, if we view their work with CWIC

in light of the principles of Lewis's 'Lexical Approach' (1997), it appears likely to be beneficial in fostering the development of a lexis-oriented view of the target language rather than a grammatically-oriented view. Lewis (1997, p.55) argued that "Are learners aware of the chunks in a text?" is a question of great pedagogic importance", and we claim to be able to answer it affirmatively for the interviewed students. The discussion above suggests that their work with CWIC was conducive to the development - or reinforcement - of two appropriate habits: borrowing pre-fabricated, attested chunks where possible rather than assembling them oneself; and attentiveness to observing with an open mind, rather than assuming, which and how many words are bonded together. We cannot claim that the CWIC work was solely responsible for developing a chunk-sensitive mentality in these students but at least that they became aware that a chunk-sensitive approach to CWIC operations brought them rewards.

Furthermore, in putting the emphasis on the process of exploring, observing and borrowing, we hope to have encouraged the students to gain a sense of achievement even from outcomes that are not totally accurate, in that they include surface errors of adaptation. We think that translating into practice Lewis's principles that "Successful language is a wider concept than accurate language" and "Language is recognised as a personal resource, not an abstract idealisation" (1993, p.vi) is useful in supporting the students' confidence and willingness to explore and borrow - highly valuable in the long-term for language learning. For example, we considered as an achievement S9's use of "Desideriamo costruirvi una vita insieme" (for 'We want to build ourselves a life together') - where she conjugated the first verb correctly for the subject 'we' but neglected to change the reflexive pronoun attached to the infinitive 'costruire', leaving 'build yourselves', as in the original (see example 2 in Table 1). We deemed this successful language in that it conveyed a precise intended meaning and was easily comprehensible. The surface error in adaptation was outweighed by the potential gain, for S9, of having chosen and adapted a pattern that a NS would use.

A second way of looking at our effective users' work with CWIC as healthy for their language learning in general is through the connection with enjoyment and emotions, suggesting a role for corpus operations in fostering and rewarding their *receptivity* to language, in van Lier's (1996) terms. In his schema of the core elements of successful language learning, van Lier (p.41) presented receptivity as essential for learners' ability to profit from exposure to language, and to move from exposure to engagement - and therefore intake and proficiency. His description of receptivity as "...related to exploration, manipulation and play", and in terms of curiosity, rather than "a passive state of openness" (p.48) closely fits the spirit our interviewees expressed. Table 4 gives some examples of ways in which students linked CWIC operations to emotions or showed enjoyment of the borrowed patterns themselves, beyond appreciation of their usefulness.

[Table 4 goes here]

Q4 In what ways does our specific approach support these learners' development as effective users of a monolingual corpus?

Our apprenticeship with this monolingual corpus takes place in a specific learning-and-teaching context and therefore, at an overall level, the effectiveness of CWIC use by our students must reflect a certain goodness of fit between those students' needs, their proficiency level and the course aims, as well as the approach taken to that apprenticeship. However, we see our approach as potentially transferrable to other contexts because it is characterised by a

set of principles and practical elements that make it a learner-centred approach and one that particularly exploits the monolingual and unmediated nature of the corpus.

The most important of these principles, as discussed above, is that we do not seek to train our students to use corpora in the ways that linguistic researchers do, but instead treat them as engaged in a completely different activity. So we do not perceive the need to address certain problems identified in other learning contexts, e.g. that “it takes a corpus linguist to offer a corpus-based writing class” (Ädel, 2010) or tackle “issues that arise due to the transfer from research context to classroom environment” (Breyer, 2011). Although our approach is still one of ‘data-driven learning’, the students are decidedly not in a learner-as-researcher role. The result of a student’s operation is typically a pattern for use in a specific sentence, chosen on the basis of understanding only one or a few examples, while a researcher tends to aim at documenting and understanding the similarities and differences between variants of a phenomenon, and defining rules that account for all relevant examples. Furthermore, judging an operation as successful even if not completely accurate, as we do for some of the students’ borrowings, is a concession few would make to a researcher.

The other key principles of our approach that we consider significant in its success are:

- We introduce the use of a corpus as a reference resource in the context of an autobiographical, creative writing task aimed at writing for a real audience of peers;
- We use a corpus and user interface appropriate to our students’ needs for that task;
- We do not try to ‘teach’ certain predetermined language points using the corpus, but to enable the students to use it independently, in ways they see fit, to enrich their writing and deal with specific lexico-grammatical issues that come up in their writing.

We believe that this approach, based on these principles, facilitated a process of authentication of CWIC use, in van Lier’s (1996) terms, by those we see as effective users. On authenticity van Lier (p.13) wrote “an action is authentic when it realises a free choice and is an expression of what a person genuinely feels and believes”. If we understand authenticity in relation to “a personal process of engagement” for learners, linked to “self determination and commitment to understanding” (van Lier 1996, p.128), then the above discussion suggests various indicators that the interviewees were authenticating CWIC use for themselves as individuals: their evident enjoyment in using it and appreciation of patterns found; their conviction regarding its usefulness; and their individualisation of CWIC use, each having developed personal habits consistent with their own learning styles. Above all, we claim that they found CWIC helped them in pursuits that *mattered* to them: finding ways to convey in greater depth or with greater feeling what they wanted to say; finding more interesting or varied ways to express themselves; and achieving greater accuracy. For example, S1 made clear that it mattered to her, for the quality of her writing, that CWIC gave NS patterns while Google might not. And some of S3’s comments conveyed her identification with the authors of CWIC texts as fellow writers, with whom she shared the concern to vary her vocabulary. Furthermore, some of the students reported using CWIC for other writing, not required for their studies: e.g. S5 for email correspondence with contacts in Italy, as she was arranging a trip there; and S15 for translating her own poems.

We see as crucial, for the students’ authentication of CWIC use, their being introduced to it first as a collection of whole texts for browsing - not just as a source of concordances. S3’s amused response to the interview question on what she had found useful in the first CWIC training session had nothing to do with training or language-learning resources: “Well, initially I think most of us found it hilarious that someone could have so many words of love

in one letter, and so many different words for ... love, like 'carissima' [dearest]..., all sorts of stuff. Mind you it's very romantic". What she recalled was clearly an 'authenticating' response - her feelings and thoughts, relating the letters to her own world (she then digressed at length comparing the CWIC letters with the love letters she herself wrote in English).

But it is also crucial that the corpus was introduced primarily as a tool to help the students carry out the autobiography task: we saw the interviewees' CWIC use as reinforcing and reinforced by their process of authenticating that task. Probably, most students initially approached the task as primarily a major item of assessment, rather than with purely intrinsic motivation. However, the way the task was presented - as an opportunity to express themselves and convey something of their own lives to their peers - and the way the process was structured during the semester (especially the use of a wiki to publish drafts and obtain feedback and reactions from fellow students) were intended to facilitate their individual authentication of it. If the use of CWIC rendered the task more rewarding, and more authentic in the sense of helping them convey more sincerely or with greater feeling what they wanted to say, then the possibility was created for the two processes of authentication - of CWIC and the writing task - to reinforce each other. A comment by S2 (*compito* 5) suggests this was true for her:

I originally typed in 'laurea' [degree] looking for a small phrase to indicate why it is good to do a degree ... but then in my search I came across this very lovely phrase 'mi piacerebbe tentare, mettermi alla prova' [I'd like to try, to put myself to the test] which completely changed what I had planned mentioning in my chapter. I feel that these few words really indicate the entire reason as to why I decided to go to uni. This whole idea of attempting to challenge myself as a way of both learning and maturing.

This comment brings together the value of the corpus and the autobiography task for S2's language-learning process: the struggle to complete a composition in Italian provided the opportunity to reflect on what being at university meant for her and to express something that mattered to her.

On the practical side, we consider that requiring the students to use CWIC for the *compiti* (or at least strongly recommending it) was fundamental to the success of the approach. Four of the interviewees made unsolicited comments indicating that this was the key to their having taken it up - that they had needed that first-hand experience to work out how to use it in ways that suited them, and they might not otherwise have bothered because it was too different from what they were accustomed to. The class forum helped in this sense too: the sharing of techniques and types of operations was remarked upon as useful by many of those present (and by S1, S2 and S3 in the interviews), for picking up new ideas from other students, not just teachers. Another practical aspect that we consider a factor in the success of the approach is the user interface to CWIC, which the students found easy and effective to work with, as noted above. And we acknowledge that we have considerable experience by now in pedagogical applications of corpora and have invested considerable time in developing the approach and evaluating it with previous cohorts of students.

Clearly, however, our apprenticeship did not convert the whole cohort into convinced and effective CWIC users, and S22's case - of limiting CWIC use to operations with very specific aims - suggests the approach did not have the full intended impact even on those who did use CWIC with successful outcomes. It is possible that the 'non-adopters' perceived the corpus work as simply an exercise in class and a step in the *compito* instructions, rather than a means to an end, and/or that they were not present for the class forum. Further investigation will be needed to understand how the approach can be improved, and particularly how we can make all students aware of the benefits of open-ended operations and borrowing.

Conclusion

In this paper we have described and sought to account for what we see as a positive experience of introducing intermediate-level students to corpus consultation, in that the seed appears to have been sown for many of them to make ongoing effective use of CWIC and other corpora. Our approach aims to exploit and highlight the monolingual and unmediated nature of our corpus, rather than presenting this as a challenge the learners must come to terms with. We present pattern-hunting as the essential function of such a resource, which therefore deserves a central place in every learner's suite of reference resources. Our case that the students discussed above are effective users of a monolingual corpus involves the kinds of successful operations they conduct, and what they do with data they find. That is, they demonstrate an observe-and-borrow mentality - of asking 'What can I use?' rather than 'What can I find out?' - along with awareness of why that approach works for them, of how CWIC relates to other reference resources, and other attitudes that support effective use. We see these as evidence of reflective, reasoned use of CWIC, likely to stand them in good stead for independent work with it as part of their regular resource kit.

While we emphasise that this approach is not oriented to transferring linguists' corpus-analytical skills to language learners, we do see it as preparing the ground for our students' later development into a learner-as-researcher role. That is, we aim to cultivate attitudes and behaviour that are essential to a researcher in any field – especially the habit of devising searches and observing data in an open-minded way in order to avoid assumptions about what will be found. At the same time, we also envisage our approach to corpus apprenticeship as likely to be effective with less experienced and/or younger learners than ours, and one of our next steps will be to trial the approach with secondary-school learners.

Indeed, although we have adopted this approach in a very specific context - to the point of using a corpus we created ourselves, tailored to our curriculum and learners' needs - we do not see it as appropriate only for intermediate-level learners or only in relation to supporting writing skills, nor purely as an introduction to work with corpora. The outcomes for our students, discussed above, suggest this approach can have far-reaching value for developing independent language learners generally.

References

- Ädel, A. (2010). Using corpora to teach academic writing: Challenges for the direct approach. In Campoy-Cubillo, M., Belles –Fortuño, B., & Gea-Valor, L. (Eds.), *Corpus-based approaches to English language teaching* (pp. 39-55). New York: Continuum.
- Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, 17(1), 47–64.
- Braun, S. (2007). Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL*, 19(3), 307–28.
- Breyer, Y. (2011). *Corpora in language teaching and learning: Potential, evaluation, challenges*. New York: Peter Lang.

Breyer, Y. (2009). Learning and teaching with corpora: Reflections by student teachers. *Computer Assisted Language Learning*, 22(2), 153-172.

Chambers, A. (2005). Integrating corpus consultation in language studies. *Language Learning & Technology*, 9(2), 111-125.

Chambers, A. (2007). Popularising corpus consultation by language learners and teachers. In Hidalgo, E., Quereda, L., & Santana, J. (Eds.), *Corpora in the foreign language classroom* (pp. 3-16). Amsterdam: Rodopi.

Chambers, A., Farr, F., & O'Riordan, S. (2011). Language teachers with corpora in mind: From starting steps to walking tall. *The Language Learning Journal*, 39(1), 85-104.

Chang, J. (2014). The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL*, 26, Special Issue 02, 243 – 259.

Frankenberg-Garcia, A. (2005). A peek into what today's language learners as researchers actually do. *International Journal of Lexicography*, 18(3), 335-355.

Frankenberg-Garcia, A. (2012). Raising teachers' awareness of corpora. *Language Teaching*, 45(4), 475 – 489.

Geluso, J. & Yamaguchi, A. (2014). Discovering formulaic language through data-driven learning: Student attitudes and efficacy. *ReCALL*, 26, Special Issue 02, 225 – 242.

Hoey, M. 2005, *Lexical priming: A new theory of words and language*. London: Routledge.

Kennedy, C., & Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning & Technology*, 5(3), 77-90.

Kennedy, C., & Miceli, T. (2002). The CWIC project: Developing and using a corpus for intermediate Italian students. In Kettemann, B., & Marko, G. (Eds.), *Teaching and learning by doing corpus analysis* (pp. 183-192). Amsterdam: Rodopi.

Kennedy, C., & Miceli, T. (2010). Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning & Technology*, 14(1), 28-44.

Kilgariff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). [The sketch engine](http://www.sketchengine.co.uk). In Williams, G., & Vessier, S. (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 105-115). Lorient, France: Université de Bretagne-Sud, Faculté des lettres et des sciences humaines. See also: <http://www.sketchengine.co.uk>

Kılıçkaya, F., (2015). Computer-based grammar instruction in an EFL context: Improving the effectiveness of teaching adverbial clauses. *Computer Assisted Language Learning*, 28(4), 325-340.

Lewis, M. (1993). *The lexical approach*. Hove, UK: Language Teaching Publications.

- Lewis, M. (1996). Implications of a lexical view of a language. In Willis J., & Willis D. (Eds.), *Challenge and change in language teaching* (pp. 10-16). Bath, UK: MacMillan Heinemann.
- Lewis, M. (1997). *Implementing the lexical approach. Putting theory into practice*. Hove, UK: Language Teaching Publications
- Mishan, F. (2004). Authenticating corpora for language learning: A problem and its resolution. *ELT Journal*, 58(3), 219-227.
- O'Sullivan, I. (2007). Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy. *ReCALL*, 19(3), 269–286.
- Pérez-Paredes, P., Sánchez-Tornel, M., Alcaraz Calero, J., & Aguado Jiménez, P. (2011). Tracking learners' actual uses of corpora: Guided vs non-guided corpus consultation. *Computer Assisted Language Learning*, 24(3), 233-253.
- Poole, B. (2011). [Review of the book *Corpus-based approaches to English language teaching* by Campoy-Cubillo, M., Bellés Fortuño, B., & Gea-Valor, M. (Eds.)] *English Language Teaching Journal*, 65(1), 92-93.
- Römer, U. (2008). Corpora and language teaching. In Lüdeling, A., & Kytö, M. (Eds.), *Corpus linguistics. An international handbook. Vol. 1* (pp. 112-130). Berlin: Mouton de Gruyter.
- Römer, U. (2010). Using general and specialized corpora in English language teaching: Past, present and future. In Campoy-Cubillo, M., Belles-Fortuño, B., & Gea-Valor, L. (Eds.), *Corpus-based approaches to English language teaching* (pp. 18-35). London: Continuum.
- Sha, G. (2010). Using Google as a super corpus to drive written language learning: A comparison with the British National Corpus, *Computer Assisted Language Learning*, 23 (5), 377-393.
- Sinclair, John (1991). *Corpus, concordance and collocation*. Oxford: Oxford University Press.
- Tono, Y., Satake, Y. & Miura, A. (2014). The effects of using corpora on revision tasks in L2 writing with coded error feedback. *ReCALL*, 26, Special Issue 02, 147 – 162.
- Tribble, C. (2015). Teaching and language corpora: Perspectives from a personal journey. In Leńko-Szymańska, A., & Boulton, A. (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 37-62). Amsterdam: John Benjamins Publishing Company.
- Van Lier, L. (1996). *Interaction in the language curriculum: Awareness, autonomy and authenticity*. London: Longman.
- Yoon, C. (2011). Concordancing in L2 writing class: An overview of research and issues. *Journal of English for Academic Purposes*, 10, 130-39.

Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, 13(4), 257–283.

Yoon, H., & Jo, J. W. (2014). Direct and indirect access to corpora: An exploratory case study comparing students' error correction and learning strategy use in L2 writing. *Language Learning & Technology* 18(1), 96–117.