# Getting the corpus habit: EAP students' long-term use of personal corpora

Maggie Charles *

Oxford University Language Centre, 12 Woodstock Road, Oxford OX2 6HT, UK

## ARTICLE INFO

## ABSTRACT

This paper reports on the long-term use of personal do-it-yourself corpora by students of EAP. Forty international graduate students attended a course in which they built and examined their own corpora of research articles in their field. One year after the course, they completed an email questionnaire, which asked about their corpus use in the 12 months since the end of the course. Results show that 70% of the respondents had used their corpus: 38% were regular users (once per week or more), 33% irregular users (once per month or seldom) and 30% non-users. Most users consulted the corpus for checking grammar and lexis while composing and revising and 93% of them considered that corpus use had improved their academic writing. Reasons for non-use included the small size of the corpus and its lack of reliability and convenience. Case studies of a user and a non-user are presented and highlight two other factors likely to affect take-up: the individual's writing process and the focus of their current writing concerns. The paper discusses the reasons behind long-term use of personal corpora and some of the challenges to be overcome in extending the approach more widely.

## 1. Introduction

The direct use of corpus data by language learners has now been a subject of study for over two decades and there has been a particular focus on the teaching and learning of academic writing in English at university level (for reviews see Boulton, 2010b; Flowerdew, 2010; Yoon, 2011). Research has been carried out with students working on three main types of corpora: large general corpora like the British National Corpus (BNC); small specialised corpora, often compiled for an individual discipline-specific course; and personal do-it-yourself corpora constructed by students themselves. One overall driver for this research effort has been to understand whether, to what extent, and under which circumstances students can usefully engage directly with corpus data as they learn academic writing. To date, the majority of the work has been based on large general corpora, but there is growing interest in the use of specialised and personal corpora.

Research with students using large general corpora lends itself particularly well to the teaching/learning of individual lexico-grammatical items and has employed both quantitative and qualitative approaches. Quantitative studies include those which used an experimental set-up to investigate whether corpus consultation had a demonstrable effect on student learning of individual items (e.g., Boulton, 2009, 2010a; Cresswell, 2007; Estling Vannestål & Lindquist, 2007). Here there was evidence that students using corpus data performed better than control groups, although results were not conclusive in all cases. Other studies, which quantified the success rate of error correction/self-correction, showed more positive outcomes, suggesting that students were able to apply corpus data to solving language problems (e.g., Gaskell & Cobb, 2004; Gilmore,

* Tel.: +44 1865 283360; fax: +44 1865 283366.
    E-mail address: maggie.charles@lang.ox.ac.uk

2009; Watson Todd, 2001). The qualitative research employed methods such as questionnaires and interviews to examine student attitudes to and evaluation of corpus use (e.g., Bernardini, 2000; Granath, 2009; Varley, 2009; Yoon, 2008; Yoon & Hirvela, 2004). Again, the results of these studies were somewhat mixed, with students' enthusiasm for corpus work often depending both on their level of English and the extent of training and support available.

Work with specialised corpora, while still addressing lexico-grammar, is also able to focus on genre and discourse issues, as discussed by several researchers (e.g., Bondi, 2001; Charles, 2007, in press; Flowerdew, 2012; Gavioli, 2005). For example, Weber's (2001) students consulted a corpus of law essays; they used concordances to identify the links between lexico-grammatical items and generic moves and succeeded in producing acceptable law essays. Improvement in writing a specific genre, research articles (RAs) in psychology, was also noted by Bianchi and Pazzaglia (2007), whose students helped compile a class corpus of relevant RAs. Hafner and Candlin (2007) provided students with a corpus of legal cases for help with their writing assignments outside class. They found roughly equal numbers of users and non-users, with the non-users divided between those who preferred other web-based tools for checking language and those who preferred not to use such tools at all. They made the point that the students had not developed long-term habits of using the corpus; rather they used it opportunistically to provide support for specific assignments.

Finally, a much smaller number of studies describes the use of personal corpora built by students themselves. Lee and Swales (2006) report on the achievements of six students who presented linguistic investigations of self-chosen aspects of their own discipline. Similarly, Gavioli's (2009) English majors wrote extended papers after compiling and examining personal corpora for their own research purposes. In both cases, the students have effectively become corpus linguists, an approach which requires large inputs of time and resources and may well be beyond reach on most academic writing courses. However, working with larger, multi-disciplinary groups and within more limited time and resource constraints, Charles (2012) also reported positive evaluations of a course in which students used personal corpora to explore discourse functions in their disciplines.

To date, however, attention has focused primarily on student achievements and evaluation of corpus work immediately after a corpus course has been completed. As Pérez-Paredes, Sánchez-Tornel, and Alcaraz Calero (2013) point out, there is little data on long-term use, and in particular on whether students carry on using corpora once their course has finished. Although Yoon (2008) tracks her students' corpus use for approximately half a year, both during and after their course, her research is based on just six case studies and thus does not provide quantitative data on continuing corpus use. If corpora are to take their place as a third reference resource alongside dictionaries and reference grammars, then it is important to determine whether student take-up persists in the longer term. The aim of the present study is to examine the extent of students' use of corpora one year after they took an academic writing course in which they constructed their own personal corpus and learnt how to use it to answer language queries.

## 2. Background to the study

This research is based on data provided by 40 international students who attended a corpus-building course in 2009 or 2010 and reported on their corpus use one year later. The course, which forms part of the Academic Writing programme at Oxford University Language Centre, is designed for advanced-level graduates and is open-access and non-assessed. Groups of approximately 15–18 students are taught in 5/6 parallel, multi-disciplinary classes. The corpus work takes place over 6 weeks, with one two-hour session per week. Classes are held in a computer lab, although some students also work on their own computers.

### 2.1. The course

The course is designed to offer participants the opportunity to create a personal corpus and practice retrieving and interpreting the information it provides. In class, students work with their own corpora in order to answer lexico-grammatical queries and explore discourse issues. Details of the programme and its rationale are discussed fully in Charles (2012); here I just summarise its major features.

The first session offers an introduction to corpora and a demonstration of the freely available concordance software AntConc (Anthony, 2007), which students use to consult a corpus of theses previously constructed by the researcher. The focus is on the potential of corpora to deal with lexico-grammatical queries. In the second session, students build a personal corpus, by selecting RAs in their fields and converting them into text files. They are also shown how to clean the files by removing matter which is not part of the running text (e.g., references, graphical elements). This procedure makes it easier to read concordance lines and renders the statistical data more reliable. Partial cleaning is recommended, but it is left to students to decide whether to clean or not and many do not do so. Each of the remaining four sessions focuses on a different corpus tool and provides detailed questions which ask students to explore a specific discourse function in their personal corpus. For example, in the session on argumentation, students use the Word List tool to explore the frequency and sentence position of linking adverbials such as *however* in their own field. By the end of the course, students are familiar with concordancing and have used the Plot tool, which gives a graphic representation of the occurrence of a search term throughout each corpus file; they have examined the words in their corpus using the Word List tool and have retrieved collocations using the Collocates and Clusters tools. Some students use their personal corpus frequently outside class and become proficient at searching and interpreting their data, but by the end of the course, all have attained a basic competence in corpus skills.