Corpus-assisted editing for doctoral students: Do-it-yourself corpora for self-correction and learning

Maggie Charles (Oxford University, UK)

From the early days of data-driven learning, one impetus for the pedagogical use of corpora arose from the need to deal with errors or infelicities in students' written work (Johns, 2002). Since then, there has been considerable interest in researching students' use of corpora for error-correction and writing improvement (Gaskell & Cobb, 2004; Gilmore, 2009; Mull & Conrad, 2013; Tono, Satake, & Miura, 2014; Watson Todd, 2001; Yoon & Jo, 2014). This research took place in a variety of pedagogical circumstances: different corpora were used, including the British National Corpus (Tono et al.) and the COBUILD Corpus and Collocations Sampler (Gilmore); corrections were carried out in class (Mull & Conrad) and independently (Watson Todd), with a time limit (Yoon & Jo) and without (Gaskell & Cobb); varying amounts of assistance were provided including pre-cast links for concordances (Gaskell & Cobb), coded error feedback (Tono et al.) and underlining errors (Gilmore). Although most research dealt with both lexical and grammatical errors, Watson Todd focused exclusively on lexical misuse, while Mull and Conrad and Tono et al. dealt only with grammatical problems. Despite their differences, all these studies reported substantial figures for accurate correction of errors at the word and sentence level, with the majority recording a success rate of 60-70%. However, most research to date has targeted undergraduate writers at an intermediate level and their writing assignments have been relatively general and non-specialist in nature. It is less clear whether students at doctoral level, with more advanced English skills and writing highly specialised texts would also benefit from consulting a corpus to deal with their errors. Thus the aim of this paper is to determine whether doctoral students working on their theses make word and sentence level errors that can be corrected using corpora and if so, whether they are able to self-correct using a tailor-made corpus and with minimal input from the teacher.

This research focuses on an EAP course designed to teach corpus-assisted editing to doctoral students. The programme consisted of six two-hour sessions in which students were introduced to the AntConc software (Anthony, 2014) and each student built two tailor-made do-it-yourself corpora (Charles, 2012, 2017; Lee & Swales, 2006). The first was a corpus of research articles (RAs) in their own field and the second was a learner corpus of the draft chapters of their own thesis. The RA corpora ranged in size from 77,000 to 3.3 million words, with a median of about 500,000 words. In class, students were shown how to use corpus tools for editing and practised both on pre-set tasks and self-chosen queries. Outside class, students also carried out a 'joint editing task' with the teacher/researcher, in which they used their RA corpus to edit a single chapter of their thesis. This paper is based on data from 20 students who did this task.

Working on an e-copy, the teacher/researcher initiated the task by using the Comment function of Word to draw students' attention to problems in their text and to ask them to use their RA corpus to deal with the issue. Problems were indicated by highlighting the relevant segment of text and stating 'Problem for corpus search' in the comment. No further hints were given as to the nature of the problem or the searches that would be appropriate. However, the teacher/researcher checked that the necessary information to amend the text was available in the student's corpus before the problem was flagged for attention. Students were asked to amend their text using the Track Changes function and to add their own comment on the problem, their search processes and results before returning the chapter. Students' amendments were checked and the chapter returned with additional comments where necessary. This paper analyses the first three problems for corpus editing in each student's text; it reports on the source of the problem, the success or otherwise of the student's edit and their comment. In two cases no correction was made and not all students commented on all problems; thus the data consist of 60 problems, 58 amendments and 52 comments.

All participants had completed a minimum of one year of doctoral studies and written at least one draft chapter of their thesis. Only one, a student of linguistics, had prior experience of using a corpus. A wide range of disciplines was covered, with half the participants studying natural sciences, seven working in social sciences and three in arts/humanities. Students spoke nine different L1s, of which the most prevalent were Chinese (8), Korean (3) and German (3).

The students' problems were analysed using the coding system developed by Chuang and Nesi (2006). This system has a hierarchical structure in which the major code refers to the language level of the error: grammatical, lexicogrammatical or lexical, while subcodes describe the linguistic category, e.g. determiner, preposition, countability of noun, or collocation. An indication of surface structure deviance is also included, consisting of five categories: omission, overinclusion, misformation, misselection and misordering. Thus a typical error tag has three parts: language level, linguistic unit and surface alteration. Example (1) indicates that a determiner-article 'the' (dtar the) has been omitted (–) and that this is a grammatical error (G).

(1) Tell-Aswad glazes are all of {dtar the – G} high-lead compositional type…

The results show that the majority of the 60 problems (58%) were at the grammatical level, while the figures for lexicogrammatical and lexical problems were similar at 22% and 20% respectively. Further breakdown of the figures reveals that the most frequent grammatical error was the use of determiners (14 instances, 23%), with omission of 'the' the single most prevalent problem, as illustrated above. The most frequently occurring lexicogrammatical problem concerned noun related preposition use (9 instances, 15%), almost all of which showed misselection of the preposition, as underlined below in example (2):

(2) There is a description <u>from</u> Tangier…

The lexical problems were more varied, although miscollocation occurred in seven instances, as seen below in the underlined misselection of the verb in example (3):

(3) …salient peripheral cues… <u>pull</u> participants' attention to that location.

These results are generally in line with those found by Chuang and Nesi (2006) for undergraduate writing. It is the ongoing persistence of such problems even at this level that renders corpus-assisted editing a valuable procedure for doctoral students.

In order to get a more detailed picture of student corpus use, both comments and corrections were analysed to determine first whether the students managed to identify the problem accurately and second whether they succeeded in correcting the problem. The results showed that most students did identify the problem correctly (93%), with only four instances of misidentification. Of course, the fact that the segment of text containing the error was highlighted certainly directed students' attention towards the problem, making accurate identification more likely. In accordance with previous research, most problems (72%) were also corrected successfully and students' comments also showed a process of induction at work. Example (4) illustrates how a Korean student investigated and explained the use of the noun 'politics':

> (4) Clusters: politics
> 44 hits for politics "is"
> 22 hits for politics "are"
> Both are possible, but I am talking about general idea by "politics" in this sentence, so it should be 'politics + singular form'.
> Amended text underlined: …how politics influences the choice of industrialization…

Although her search technique can be criticised, the student managed not only to deal with the problem, but also to formulate a guideline for her own future use.

Of the 17 amendments classified as unsuccessful, six were due to unsatisfactory searches and four to misidentification of problems; there were three instances of partial success, two without any correction and one instance each of an erroneous explanation, and a correction without corpus use. However, even unsuccessful attempts at editing can promote the noticing of useful language, as shown in this comment (5) by a Chinese student:

> (5) I launched a search of "echo*" and got 19 hits. After sorting them with the logic of 1R, 2R and 3R, I noticed several collocations, namely 1) X is echoed by; 2) X echoes in/ has echoes in. So I would replace "echoing with" with "echoing in".
> Unsuccessfully amended text underlined: Echoing in the call for community level self-governance…

Although the student did not notice instances of the verb 'echo' with a direct object and thus did not correct the problem successfully, he did identify other important patterns associated with the verb. Moreover, his comment shows that he can make good use of search techniques and notice frequently recurring patterns in the data. Despite his lack of success on this occasion, then, he is well on the way to becoming competent at corpus-assisted editing.

This paper argues that even students at advanced doctoral level can benefit from the use of corpora for self-correction. Not only do students still make word and sentence level errors that they can correct using their corpus, but in the process they are able to induce rules and make discoveries that enhance their disciplinary and linguistic awareness.

References

Anthony, L., (2014). AntConc (3.4.4). [computer program] Tokyo, Japan: Waseda University. Available at: <http://www.laurenceanthony.net/>

Charles, M. (2012). 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes*, *31*(2), 93–102.

Charles, M. (2017). Do-it-yourself corpora in the classroom: Views of students and teachers. In K. Hyland & L. Wong, (Eds.), *Faces of English education: Students, teachers and pedagogy* (pp. 107–123). Abingdon: Routledge.

Chuang, F.-Y., & Nesi, H. (2006). An analysis of formal errors in a corpus of L2 English produced by Chinese students. *Corpora*, *1*(2), 251–271.

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, *32*, 301–319.

Gilmore, A. (2008). Using online corpora to develop students' writing skills. *ELT Journal*, *63*(4), 363–372.

Johns, T. (2002). Data-driven learning: The perpetual challenge. In B. Ketteman & G. Marko (Eds.), *Teaching and Learning by Doing Corpus Analysis* (pp. 107–117). Amsterdam: Rodopi.

Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, *25*(1), 56–75.

Mull, J., & Conrad, S. (2013). Student use of concordancers for grammar error correction. *ORTESOL Journal*, *30*, 5–14.

Tono, Y., Satake, Y., & Miura, A. (2014). The effects of using corpora on revision tasks in L2 writing with coded error feedback. *ReCALL*, *26*, 147–162.

Watson Todd, R. (2001). Induction from self-selected concordances and self-correction. *System*, *29*, 91–102.

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 and Technology*, *18*(1), 96–117.