Using a Corpus as a Resource for Writing

Information for Supervisors

Maggie Charles
maggie.charles@lang.ox.ac.uk

1 Introduction: Graduate Writing Issues

Graduate students may experience a number of problems with their writing which could be tackled by using a corpus together with text analysis software.

There may be **sentence-level issues of grammar and vocabulary choice**. For example, the writing may not sound 'natural', '*a*' and '*the*' may be omitted before nouns or the choice of preposition may obscure the meaning.

There may also be **text-level issues concerning the development of the content**. For example, students may be uncertain how to deal with the literature in their field or they may wonder how their results should be discussed.

Using a corpus can provide information to help students with issues like these.

2 Corpus and Concordance

A corpus (plural: *corpora*) is a collection of electronic texts. Corpora are built with a specific purpose in mind and are designed according to specific criteria. For example, if the student's aim is to check the language used in a specific field, a suitable corpus could consist of relevant research articles written by experts. Corpora are accessed by using **text analysis software.**

Text analysis software analyses the texts in the corpus automatically and presents the results on screen. It provides several different tools for examining the data, the most widely-used being the **concordancer**.

The concordancer searches the corpus for every instance of a word or phrase specified by the user and presents each one with its context in a line on screen.

The **search term** appears in the centre, usually with about 5 or 6 words either side of it. This is called a 'keyword in context' or **kwic** concordance.

For example, the phrase '*a broad variety of…*', which appeared in a student's writing, sounds rather unnatural. Example 1 gives part of a concordance on the search term *variety*, retrieved from a 500,000 word corpus of Oxford theses. The concordance shows that the most frequently used phrase is '*a **wide** variety of…*'. The adjective *broad* does not occur at all with the noun *variety* in this corpus. Words often occur in combinations which are more or less fixed, that is they **collocate** with other words. Using a corpus and concordance can help students find information on these collocations, which is often not available in a dictionary or reference grammar.

Example 1: A Concordance on the Word '*variety*'

```
1           s that the carbon black in one batch of a given variety (e.g. Raven 430) has a large
2             strong 1993: 13-15). <ptr3> Although the historical variety of ideas of international
3       e neatness of a cycle imposed by Vico on the huge variety of human history - then we are
4           the severity. With these qualifications a large variety of samples have been measured
5           ommensurable and all contributing to the rich variety that makes up Humanity. Mazzini
6             s</head>   These materials exhibit the richest variety of chemistry including accept
7         arkises wicked by-blow?"'(GR 171. Here the standard variety of English at the beginning
8             s demonstrated that it depends   upon a wide variety of factors. This has enabled the
9         ools for future studies of treatments on a wide variety of materials. The details of the
10           ll ate widely differing diets. Yet with a wide variety of foods available to north west
11     the samples measured would have resulted in a wide variety of decay forms and decay rates
12       the biosynthesis of amino acids occurs via a wide variety of metabolic routes, the carbon
13       e of taking several thousand STM images of a wide variety of metallic surfaces [3.15].
14       asured at raised temperature (160(C), from a wide variety of natural samples, could be 23
15         43 0.5. As shown in Figures 7.4 and 7.5, a wide variety of shine-plot forms are possible
16             pe was used<ptr3>.   <p>Because of the wide variety of electron microscopes availabl
17           oduces gram quantities of material, the wide variety of conditions within the confine
18       ining the dose response characteristics of a wider variety of samples;   (vii) the
19       wide range of resolved emission bands and a wider variety of storage times/temperatures.
20       of n (to perhaps exponential or Gaussian) a wider variety of decay forms could be
```

Students can build two types of corpora that may be useful for their writing.

1. **A corpus of expert texts**, e.g. research articles in their field
   - Useful because the language is specific to the field and relevant to their own writing needs.
2. **A corpus of their own writing,** e.g. chapters of their thesis or their research papers
   - Useful because students can examine their own writing automatically to help identify problems.

Further details on how to build corpora and use the software are available in the pages for students: 'Using a Corpus as a Resource for Writing'.

3 Further Information

3.1 AntConc Concordance Software

There are many concordancers available, but students are recommended to use a freeware programme called *AntConc* (version 3.2.4) produced by Professor Laurence Antony at Waseda University, Japan. This software has been in use at the Language Centre for several years and has proved reliable and easy for students to operate.

The concordancer is the tool most often used by students and is described in detail in the student pages: 'Using a Corpus as a Resource for Writing', but *AntConc* also provides several other tools. Further information on all aspects of the software can be found at Laurence Antony's website here: http://www.antlab.sci.waseda.ac.jp/antconc_index.html

3.2 Courses Using Corpora at the Language Centre

There are two types of course using corpora and *AntConc* run by the Language Centre. Both are for non-native speakers of English only.

1. For DPhil Students only: *Edit your Thesis with Corpora*

This course is designed to help students who are writing up their thesis. It is a 6-week (12 hour) course that runs every term starting in Week 3.

The syllabus is available here:

http://www.lang.ox.ac.uk/files/Syllabi/Doctoral%20Training%20Syllabus%202013.doc

2. For all Students: *Writing in your Field with Corpora*

This course is designed to help students examine the ways in which writers in their field use language to perform academic functions such as making arguments and citing other researchers. It is part of the Academic Writing Programme and runs in Trinity Term only.

The syllabus is available on the Language Centre website here:

http://www.lang.ox.ac.uk/courses/english.html

3.3 Introduction to Corpus Linguistics Course

The IT Learning Programme offers an introductory course in corpus linguistics, which consists of 6 lunch-time sessions. The details are here:

https://weblearn.ox.ac.uk/portal/hierarchy/central/oucs/itlp_courses/corpus_ling

3.4 Research on Using Corpora with Students

The use of corpora with students has been extensively researched over the last decade. Two recent publications deal specifically with graduates who have built their own corpora for language learning purposes at Oxford.

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. 2014. Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes* 35: 30-40.

Other References

Flowerdew, L. 2012. *Corpora and Language Education*. Basingstoke, UK: Palgrave Macmillan.

Frankenberg-Garcia, A., Flowerdew, L., & Aston, G. (Eds.). 2011. *New Trends in Corpora and Language Learning*. London: Continuum.

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.