An introduction to working with written and spoken corpus data

Chris Tribble
King's College, London
christopher.tribble@kcl.ac.uk

Guy Aston
University of Bologna, Italy
guy@sslmit.unibo.it

1 Overview

Corpus applications in language education are often associated with large scale corpus projects such as the British National Corpus (2001), or the Corpus of Contemporary American English (Davies, 2010). However, while these large corpora have been invaluable for the elaboration of lexicographic and grammatical accounts of language, they have been found problematic for many language learning and language teaching applications as they often provide either too much and too complex material, or they offer too little that is relevant to the needs of specific groups of learners.

A response to this concern can be found in the development of small or specialist corpora (Tribble, 1997; Ghadessy & Roseberry, 2001; Nesi & Garner, 2012), and their exploitation for pedagogic purposes. Through the analysis of such small corpora, it is possible for teachers to begin to develop curriculum specifications for ESP/EAP courses, and to develop supplementary materials to support learners on specialist programmes or on general programmes where there is a need to support the expansion of students' knowledge of and ability to use the grammar and lexis of a language.

In this workshop, you will have the opportunity to develop your own pedagogic corpus and to develop learning / teaching materials for classroom purposes. No previous experience of classroom applications of corpora is required, but it will be important to bring with you an idea of the kinds of students you wish to support, and, if possible, to bring a collection of texts which can be worked on during the session

2 Requirements

TEXTS

Participants should, ideally, bring with them a collection of electronic texts which can be used as a micro corpus. These might include:

- examples of student writing
- collections of specialist texts (e.g. research articles, administrative documents, informational documents etc.)
- print journalism
- fiction texts
- recorded speech (audio or video) with transcripts

If you are not able to bring a collection yourself we will be able to provide a collection of UK and US journalism (good for advanced general English learners, a Fiction collection (surprisingly good for intermediate learners), and a collection of science and social science research articles (good for advanced EAP). We will also provide a collection of audio material which could serve as a basis for work on advanced spoken/listening skills.

USB DRIVE

Participants should bring their own USB drive (at least 2G available storage)

3 Resources

We will provide learners with a Windows computer with Wordsmith Tools v6 (commercial software) and AntConc (freely available) installed, along with basic Office applications (Word / Excel).

4 Outcomes

By the end of the 3 hour workshop, participants will be able to generate wordlists, ngram lists and edited concordances which can be used as the basis for classroom materials.

References

The British National Corpus, version 2 (BNC World). (2001). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Davies, Mark (2010). "The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English". *Literary and Linguistic Computing* 25 (4): 447–65

Ghadessy, M., Henry, A. and R. Roseberry (eds.), (2001). *Small corpus studies and ELT*. Amsterdam / Philadelphia: John Benjamins.

Nesi, H. & S. Gardner (2012). *Genres across the disciplines: student writing in Higher Education* . Cambridge: Cambridge University Press.

Tribble, C., (1997). Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching.In Melia, J. and B. Lewandowska-Tomaszczyk (ed.) *PALC 97: practical applications in language corpora*, Lodz: Lodz University Press.