Using lecture slides to create an academic corpus

Simon Smith *Coventry University, UK*

The idea of supporting language learning with the use of corpora has been around since at least 1991, when Tim Johns coined the term *data-driven learning* (DDL). This approach invites learners to tease out patterns from authentic text, and test their own linguistic hypotheses in the manner of a mini research project; it has an intuitive appeal to teachers who favour student-centred or inductive learning.

In this study, students constructed and consulted their own web corpora, based on presentation slides from their subject lectures. This was intended to help them build up their subject vocabulary, as well as access authentic texts from their discipline. Although no firm conclusions about the success of the method were reached, and no formal evaluation was conducted, feedback gained from participants was positive.

The corpus query tool Sketch Engine (http://sketchengine.co.uk/) was used for the study. It includes a web corpus generating module, WebBootCat. Readers may also be interested in the freely available version of BootCat, at http://bootcat.sslmit.unibo.it/.

Corpus construction by learners: background

Many approaches to DDL involve the *consultation* of corpus resources. It has been claimed that corpus *construction* by learners, followed by consultation, may yield better learning outcomes (Aston 2002). Involvement with the process of corpus construction may instil in learners a sense of ownership of the product, as well as leading to the acquisition of transferable skills (in ICT, for example: see Boulton 2011). Smith (2011) had language learners in a Taiwan university create their own personal corpora based on keywords.

This study

The present study differs from Smith (2011) in that (1) the corpora created were focused on the students' subject area, Accounting and Finance for International Business (AFIB), not selected by students; and (2) the corpora were seeded from course materials (mainly lecture PowerPoints) in the subject area, and not from user-selected (and sometimes arbitrary) keywords.

A group of six AFIB students undertook the corpus construction as part of an in-sessional English for Academic Purposes (EAP) class, over a period of four teaching weeks. In the first two lessons, an introduction to using corpora and reading concordance lines was given. In weeks 3 and 4, students constructed and consulted their own corpora.

Corpus construction method

The corpus construction architecture is shown in Figure 1. Corpus construction is seeded with a set of user-supplied keywords: first a search engine module finds web pages that are "about" the keywords, then other BootCat software components extract text from the web pages and generate the corpus, which can then be consulted in various ways.

First, the user uploads the text content of one or more PowerPoint (or certain other format) files to form a mini-corpus, using the Sketch Engine Corpus Architect. Because of the nature of lecture slides, the resulting corpus does not contain many full sentences; however it will include key topic vocabulary. Students could opt to create a more specialized corpus, consisting of perhaps just one PowerPoint, for example on "Capital Investment Appraisal", to which two lectures were devoted. Alternatively they might decide to create a broader corpus, such as "Management Accounting for Business Decisions", by downloading all PowerPoints on the module.

The Sketch Engine generates a list of the most salient words in the corpus (words found much more frequently in this corpus than in a standard [reference] corpus).

Using the wordlist, BootCat then creates a larger corpus, consisting of texts from the web.

Corpus consultation

The large corpus can be used to

1. produce lists of subject area words and terms for study
2. view word sketches, which give a one-page view of the collocations and grammatical structures in which a word or term participates.
3. view the words and terms in context, using concordancing
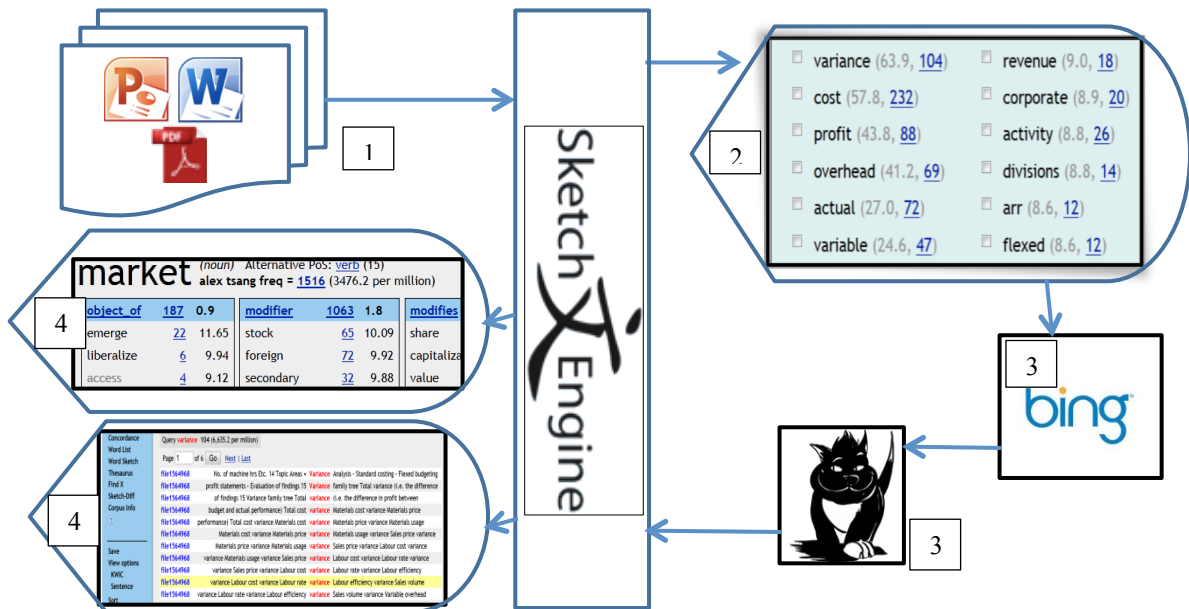4. link to the original texts on the web.

Figure 1 Schematic of corpus construction and consultation. Key: 1. Text input. 2. Wordlist from mini-corpus. 3. Bing API interacts with BootCat. 4. Word sketch and concordance displays from web corpus.

Feedback

Representative comments included:

> *The work is useful for my AFIB study. Because the software list the word which I do not understand very clearly. I can learn the speech and the meaning of this word.*

Future plans

In the coming academic year, a more detailed quantitative study will be conducted using control and experimental groups. As new lecture slides become available week by week, the learners will be encouraged to update their personal corpora, and the impact of the approach on learning will be assessed by pre/post-tests.

simon.smith@coventry.ac.uk

References

Aston, G. (2002). 'The learner as corpus designer' in B. Kettemann and G. Marko (eds.). *Teaching and learning by doing corpus analysis*. Amsterdam: Rodopi. (Retrieved 24 April 2014 from http://www.sslmit.unibo.it/~guy/graz.htm)

Boulton, A. 2011. 'Bringing corpora to the masses: Free and easy tools for language learning' in N. Kübler (ed.). *Corpora, language, teaching, and resources: From theory to practice*. Bern: Peter Lang. (Retrieved 24 April 2014 from http://hal.archives-ouvertes.fr/docs/00/32/69/80/PDF/XXXX_boulton_TaLC_interdisciplinary.pdf)

Smith, S. 2011. 'Learner Construction of Corpora for General English in Taiwan'. *Computer Assisted Language Learning* 24(4): 291-316. (Retrieved 24 April 2014 from http://curve.coventry.ac.uk/open/items/96362d30-9307-404e-81e8-16091fb2c118/1/)