DIY Corpora for Vocabulary Learning

◎ Simon Smith

INTRODUCTION

One of the principal applications of corpora in English language teaching and learning has been the compilation of vocabulary lists for student use. West' s General Service List (GSL; 1953) was based on a painstaking (manual) corpus analysis of frequency and range (Gilner, 2011) , and almost all subsequent lists, whether of general English (College Entrance Examination Center, 2002) , academic English (Coxhead, 2000) or specialist domains have been derived directly or indirectly from corpora. Learners need to acquire words that are both frequent in the language and occur across a range of texts, and the use of corpora can furnish lists that satisfy these frequency and distributional requirements.

There is a core English vocabulary which dominates many genres and styles, and it is of course important for learners to acquire this vocabulary. The General Service List, even decades after it was compiled, was found to cover 90–92% of tokens in three children' s fiction texts (Hirsh and Nation, 1992) , and 76% of tokens in the Academic Corpus, used by Coxhead (2000) to create the Academic Word List (AWL) . This list, in its turn, is intended by Coxhead to represent a core "academic" vocabulary, and forms the basis for a host of academic vocabulary activities, textbooks and learning websites, as well as inspiring other academic wordlists that followed.

In many professional and academic contexts, however, learners wish to acquire the vocabulary and terminology of their own specialist domain, which by its nature will not emerge as salient in a general corpus or appear on a wordlist derived from the same. A great deal of prior work has been done on the construction of corpora in specialist domains, and the compilation of wordlists based on them; some of this work will be surveyed in the Literature Review section. In that section, I will also consider wordlists that incorporate multi-word units (MWUs) , which are of importance in the acquisition of specialist language.

The present paper proposes a data-driven learning（DDL）approach to the creation of specialist vocabulary lists and terminological resources. University students whose first language is not English are asked to construct a corpus from learning materials and texts supplied by their specialist subject tutors. They then expand the corpus, using software tools provided, to add related texts from the Web. Next, they generate a list of the salient words and MWUs from their extended specialist corpus. Finally, they incorporate selected words and terms from the lists into their own personalized vocabulary portfolio, where they also include definitions, corpus/dictionary examples, and any other information they wish to record. The portfolio is in a spreadsheet format which they can conveniently consult and add to throughout their course（and indeed into the future）.

The following research questions will be addressed:

1.How effective are corpus construction and the compilation of vocabulary portfolios by learners in the acquisition of specialist terminology?

2.What are learners' perceptions on learning vocabulary via corpus construction and vocabulary portfolios?

These questions are addressed by means of（i）pre- and post-tests which attempt to discover to what extent the interventions helped learners in the acquisition of domain-specific vocabulary, and（ii）a questionnaire-based analysis of the perceptions of the learners about the approach.

Organization of the Paper

The next section sketches the DDL approach and relevant prior work, attempting to show how language in general（and vocabulary in particular）is more likely to be retained when the learner engages with the corpus and portfolio construction, making decisions about selection and inclusion along the way. I also look at the ways corpora have been used to create wordlists and vocabulary resources.

In the Methodology section, I first summarize a pilot study（Author, 2015）in which a small number of students constructed their own corpora and investigated concordances and collocational patterns, but were not asked to create vocabulary portfolios.

I then give further details of the main intervention reported in this paper, as well as the pre- and post-test procedures that were used to establish its effectiveness, and the qualitative perceptions study. A Results and Discussion section will present and analyse the findings from these tests, as well as conclusions drawn from participant questionnaire responses. Limitations of the study, conclusions, and directions for future research will then be presented.

LITERATURE REVIEW

In this section, I first present some of the literature on corpus consultation by learners, and the pros and cons of such a DDL approach. I then look at the ways corpora have been used professionally to produce wordlists. A third and final subsection describes prior work on corpus *construction* (rather than simple *consultation*) by learners, paving the way for work on DIY corpus-based vocabulary resources.

Background to DDL

The use of linguistic corpora in language learning often takes the form of concordance analysis by students, or data-driven learning (DDL) . In a parallel to data-driven computational algorithms, DDL attempts to impart linguistic knowledge by making available samples of authentic language, from corpora, and inviting language learners to discover usage patterns for themselves. The 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. Johns (1991) , who coined the term, likens the language learner (on the DDL model) to a researcher, analysing target language data and becoming familiar with the language through the regularities and consistencies encountered. Johns (1991: 2) , famously, goes on to claim that "research is too serious to be left to the researchers" .

What is important to note about this use of corpora in language learning is that the data are *authentic* (because a corpus contains examples of real language in use, as opposed to the possibly inauthentic examples in a textbook) , and that they are *representative* (because a corpus of, say, billions of running words, will offer plenty of examples, while a dictionary might only have a few) . As Stubbs (2002: 221) has it, corpus linguistics is both "inherently sociolinguistic" , in that the data are authentic, and "inherently quantitative [···] mak[ing] visible patterns which were only, if at all, dimly suspected" .

An early and often cited set of DDL materials is Johns's kibbitzers, of which an excellent example is presented in the 1991 paper. The title of the paper is 'Should you be persuaded' , for the reason that it presents first an activity which challenges the reader to identify (from concordance data) the several senses of the word "should" , and then another activity which invites us to characterize the difference between "persuade" and "convince" , again by appealing to supplied corpus evidence. Johns's kibbitzers (to be found at http://

www.lexically.net/TimJohns/) inspired the MICASE kibbitzers, the work of John Swales and colleagues (Regents of the University of Michigan 2011) , archived at https://web.archive.org/web/20111008033810/http://micase.elicorpora.info/micase-kibbitzers. A number of other websites and books, including Tribble & Jones (1990) , the now out-of-print Thurstun & Candlin (1997) , Reppen (2010) , and Lamy & Klarskov Mortensen (2012) offer suggestions for DDL tasks. A collection of DDL resources has been gathered by Neufeld (2012) at http://www.scoop.it/t/data-driven-language-learning/.

DDL has not, however, become widely accepted as a language teaching approach. Boulton (2008) considers a number of reasons as to why this should be so, concluding that "In a nutshell, learners and teachers simply aren't convinced" . It is the case, too, that in its default and rather prosaic consultation mode, DDL can consist of entering keyword queries at a computer keyboard and reading through lines of concordance output (or reading printed lines) . As Kilgarriff et al. (2008) put it, "The bald fact is that reading concordances is too tough for most learners. Reading concordances is an advanced linguistic skill."

Students new to corpus studies are sometimes uncomfortable with the alarming physical appearance of KWIC concordances (Lamy & Klarskov Mortensen, 2007) . Boulton (2009) summarizes what others have said about the problem of learning from truncated sentences in KWIC output, citing on the one hand Johns (1986:157) who claims that learners are quick to "overcome this first aversion" , and on the other hand Yoon and Hirvela (2004: 270) , who report that 62% of their students perceive sentence truncation as a "difficulty" .

Tim Johns' s (1991) idea that language learning should be based on research is echoed by Bernardini (2000) , who treats DDL as a voyage of discovery, serendipitous in nature, where the learner may be sidetracked along the way. Lee & Swales (2006) characterize the approach, less glowingly, as *incidentalism* (whilst admitting to having adopted it in their own study) . Whilst supporting the approach, Adel (2010: 46) , in an article on the use of corpora to teach writing, claims that students can be overwhelmed by the sheer amount of data available, and that "teacher-guided settings and clearly defined tasks" help them out of the "maze" . In her development of the work of Johns, Gavioli (2009: 47) suggests that in order to allay potential *embarras de choix*, "autonomy needs to be guided and educated" . Vincent (2013) also refers to the desirability of taking a guided discovery, rather than a purely inductive and serendipitous, approach—particularly with students new to DDL.

Gavioli also notes (p. 44) that students are particularly motivated by working with their own corpora, and that "creating and analysing corpora is something that students may take very seriously" . The students in the present study were, as has been noted, tasked with constructing their own corpora, and developing wordlists based on them. I will briefly survey prior work on learner corpus construction in the final subsection of this literature review. Next, however, I look at corpus-informed wordlists and the surrounding literature.

Corpora and Academic Wordlists

The General Service List（GSL）, as is evident from its name, lists vocabulary that was（in 1953）in general use, and does not specifically target academic needs. Through the 60s and 70s, several new academic wordlists emerged. These wordlists were generally compiled by teachers, without the aid of computers, to meet specific local needs, and were based on corpora of textbooks and other academic writings; these include Campion & Elley（1971）, Praninskas（1972）, Lynn（1973）, Ghadessy（1979）. In 1984, Xue & Nation combined the four most recent of these lists to form the University Word List. Coxhead（2000）perceived the need for an academic wordlist based on a larger corpus and more principled inclusion criteria, and her well-known AWL was generated from a 3.5m word Academic Corpus. The words admitted to the list were subject to specialized occurrence（not in the GSL）and range（cross-disciplinary reach）criteria, and were required to occur at least 100 times in the Academic Corpus. The list is organized by word family, not by word token or lemma. Thus, 'introduction' and 'argumentation', which one might expect to find on a list of academic words, are both excluded because 'introduce' and 'argue' exist in the GSL in non-academic senses.

The AWL is widely known in the academic English teaching profession, and there are a number of coursebooks and English learning websites that exploit it as an inventory of academic vocabulary. Other general academic wordlists have since been established, chief among which are the New Academic Word List（NAWL, based on the Cambridge English Corpus; Browne, Culligan, & Phillips, 2013）, and the Academic Vocabulary List（AVL, based on the COCA corpus; Gardner & Davies, 2013）. Some learning materials have been developed around the AVL, mainly on the compilers' websites vocabulary.info and wordandphrase.info, but they are not as extensive as those of the AWL.

The new lists, unlike AWL, do not conflate all derived forms into one word family. Research findings（e.g. Schmitt and Zimmerman, 2002: 158）indicate that the acquisition of one member of a word family does not necessarily facilitate the acquisition of a second member, as with the examples of 'argue' and 'introduce' noted above; or the problematic inclusion of both 'briefed' and 'brevity' under the AWL headword 'brief', where two entirely different word senses are involved.

The corpora used to compile the academic wordlists are partitioned by academic discipline: AWL is divided into four overarching disciplinary sections（Arts, Commerce, Law, Science）, each of which is further subdivided into 7 subject areas. No attempt is made to assign the words themselves to disciplines, however. Hyland & Tse (2007) point out that some senses of words (and indeed certain derived forms within AWL word families) are more likely to occur in one discipline than another. Thus, for example, the form ‘appendixes’ probably only occurs in biological or medical writing, while other members of the ‘appendix’ word family, including the alternative plural ‘appendices’ , will occur in many disciplines. The senses of other AWL word families, for example ‘revolution’ , are entirely different in say politics and engineering, but the wordlist offers no way to tease the senses apart.

Hyland & Tse (2007) investigated the distribution of AWL words in their own academic corpus, and found considerable variation in the ways words are used across the disciplines. For example, ‘process’ was far more likely to act as a noun in the sciences, with nominalization being more common there generally. Members of the word family ‘analyse’ are used differently across disciplines, often participating in highly domain-specific multi-word forms such as ‘genre analysis’ and ‘neutron activation analysis’ . Hyland & Tse (2007: 247) conclude that “A growing body of research suggests that the discourses of the academy do not form an undifferentiated, unitary mass, as might be inferred from such general lists as the AWL, but constitute a variety of subject specific literacies.” In line with Hyland & Tse’ s arguments, a number of discipline-specific academic wordlists have emerged. For example, the Medical Academic Word List (Wang et al., 2008) is based on a corpus of medical research articles; the Engineering Wordlist, the work of Mudraya (2006) , comes from engineering textbooks.

Like AWL, NAWL and AVL, these specialized lists do not include multi-word units (MWUs) . There are at least two lists of academic MWUs: the Academic Formulas List (AFL; Simpson-Vlach & Ellis, 2010) , and the Phrasal Expressions List (Martinez & Schmitt, 2012) . These, however, contain general academic MWUs, rather than discipline-specific terms. This leaves ESAP practitioners with little to go on in terms of discipline-specific MWUs.

The present research addresses these issues in that learners were encouraged to include both single word and multi-word items in the wordlists (vocabulary portfolios) they created, specific to their own discipline.

Students select texts and websites related to their specialism or area of interest to generate the corpus and portfolio, and need to make decisions about what to include, so the task is an authentic one in terms of the TBL characteristics noted by Van den Branden (2013:629) —much of the language studied by learners is acquired from authentic sources such as learning materials supplied by subject tutors. Learners acquire language—both the terminology of their subject, and the contexts in which the terms are used—as their research proceeds; they truly “learn language by using it” , in Van den Branden’ s words.

Construction of Corpora by Learners

We now return to data-driven learning. The approaches to DDL described in the first subsection of this literature review involve the *consultation* of corpus resources. It has been claimed that corpus *construction* by learners, followed by consultation, may afford better learning opportunities (Aston 2002) . The process of creating a corpus, according to Tyne (2009) inculcates a sense of ownership in the learner and therefore has a motivational impetus, and Lee & Swales (2006) emphasized this "ownership" in an apparently successful bid to get their students to engage with corpus construction, despite the students' initial reluctance. Zanettin (2002) had learners compile a corpus from the web, and analyse it with Wordsmith Tools, reflecting (p. 7) that "constructing the corpus was as useful as generating concordances from it" . Charles likewise highlights (2012: 101) the "truly revelatory moment when they see the patterns appear before their eyes*in their own data*" [emphasis in original].

Moreover, the process of compiling the corpus may lead to the acquisition of not only language, but also useful transferable skills, including IT and problem-solving competencies (Boulton 2008; Jackson 1997) . Once the corpus is constructed, some students may be sufficiently motivated to consult it and add to it when needed (Charles 2014) . Lee & Swales (2006) report that some of their students even purchased their own copies of Wordsmith Tools, indicating a commitment to continuing with corpus construction and analysis in the future.

Castagnoli (2006) had translation trainees use the BootCaT toolkit (Baroni & Bernardini, 2004) to generate web corpora on specific topics, and extract lists of terms, which could be used to compile glossaries and term databases. The students found that a larger number of relevant terms could be extracted when the domain chosen was highly specialized. By way of assessment, the students were given a technical translation task, and were asked to prepare for it by building a web corpus in the relevant domain, and extracting from it a glossary of terms.

Author (2011) extended Castagnoli's approach to non-specialist language learners in a Taiwan university. Corpus construction was seeded or bootstrapped from a set of user-supplied keywords: first a search engine module found web pages which were "about" the keywords, then other BootCat software components extracted text from the web pages and generated the corpus. Students were asked to construct and consult a corpus relating to their own academic discipline, and provide analysis and commentary, with one student, for example, commenting:

Creating a specialized corpus could be useful when it comes to researching a particular subject or learning a subject in English. It is useful because of the different results which are much more relevant than searching on a much more general English corpus.

METHODOLOGY

Pilot Study

A group of six Accounting and Finance for International Business（AFIB）students undertook a corpus construction task as part of an in-sessional English for Academic Purposes（EAP）class. The students were final year direct entry international students, having completed the first two years of their course at an institution in their home country（in these six cases, China）. An IELTS score of 6.5 is required to enter the year, and all were at this level.

The study was conducted over a period of four teaching weeks, and is reported in greater detail by Author（2015）. In the first two lessons, an introduction to the use of corpora and the reading of concordance lines was given. In weeks 3 and 4, students constructed and consulted their own corpora, based on texts and learning materials that had been made available by their AFIB module tutors. They were not asked to make vocabulary portfolios, as with the present study, but they did study concordances and consult Word Sketches（one-page summaries of word usage）in the Sketch Engine corpus analysis tool（Kilgarriff et al 2004）, focusing on academic and accounting words and terms from their corpora.

The students were asked（at the end of a homework task sheet）whether the Sketch Engine was useful for（Q1）English study and/or for（Q2）AFIB study, and（Q3）whether they found the work interesting. Two of the six students responded, both making only positive comments: the approach was useful for EAP and AFIB study, and Student 1 commented that it was "interesting and amazing". Student 2 wrote that "the process of create my own corpora was very enjoyable and makes me sense of accomplishment", confirming the findings of others reported in the Literature Review section that the process can be motivating and engaging.

Despite the indication of satisfaction, it seemed to the researcher that the students needed more of a sense of purpose when consulting their DIY corpora. They seemed quite content to explore the corpora in a more or less serendipitous way, but like Adel（2010）and Vincent（2013）, I felt that the discovery process required more clearly articulated tasks and learning outcomes. The requirement in the main study to create vocabulary portfolios met that need, as well as providing a useful reference resource for students.

Main Study

Participants

This study constituted a larger scale, quantitative follow-up to Author（2015）, and was run over a period of one (11-week) second semester. The entire cohort of AFIB top-up students (n=94) , consisting of 4 EAP class groups, participated in the study. With the exception of either one or two members of each class, all were L1 Chinese speakers; in other respects, the composition was the same as for the pilot study. Two of the class groups (EFA3 and EFA4) acted as control groups, and two (EFA1 and EFA2) as experimental groups, these last being taught by the researcher. The experimental group classes were conducted in a computer lab, and in addition to the normal EAP work specified by the syllabus, students were given the opportunity to do corpus-based vocabulary work, as described in the Intervention section, for an average of 20 minutes per two-hour weekly class.

The control groups were each week given a list of financial domain vocabulary to study in their own time. The lists were generated from Accounting & Finance corpora, created by the researcher in the same way as the students in the experimental groups created their own vocabulary portfolios (as described in *Corpus Construction* below) .

Two sub-domains of vocabulary were studied, related to two of the financial modules that all participants were studying in their home department. EFA1 (experimental group) and EFA3 (control) focused on Management Accounting (ACC) , while EFA2 and EFA4 explored the vocabulary of International Finance (FIN) . ACC and FIN are two of the three content modules followed by AFIB students in the second semester.

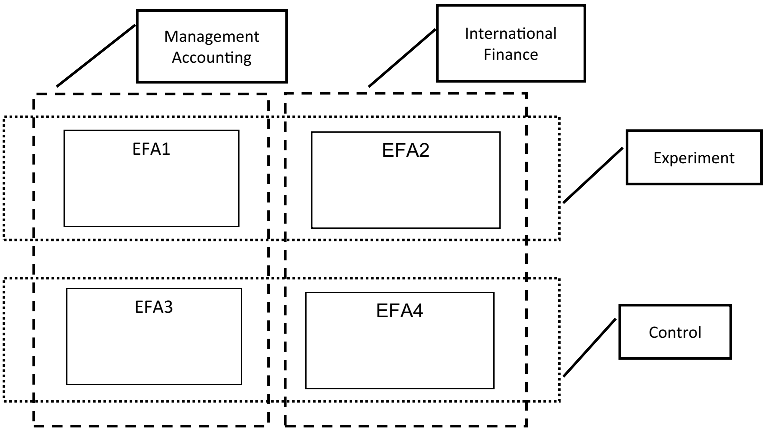


Figure 1 Configuration of participant groups

The participant configuration is shown in Figure 1. It was predicted that

H1.All groups will perform better in the post-test than they did in the pre-test.

H2.EFA1 and EFA2 will improve more overall in the post-test than EFA3 and EFA4.

H3.EFA1 will improve more than EFA2, and EFA3 more than EFA4, on Management Accounting items.

H4.EFA2 will improve more than EFA1, and EFA4 more than EFA3, on the International Finance test items.

It would have been possible to configure the groupings in a simpler way. For example, all groups could have been exposed to both of the domain vocabularies, with two experimental group classes creating corpora and the two classes using lists. This would have entailed, however, that the vocabulary being acquired from combined module resources would not have represented a coherent domain; I wanted the students to benefit from vocabulary resources that aligned with a plausible subject of study (one of their content modules) and were perceived as such.

Pre– and Post–tests

A pre–test, designed to test participants' knowledge of vocabulary in both the financial sub–domains, was administered at the start of the semester. The test included abbreviation items such as NPV and IMF, which the participants were asked to expand (in this case, to "net present value" and "International Monetary Fund" respectively) . There were also 10 gap–fill questions, for example "The bonds are trading at only 40% of f____ v____" , to which the correct answer would have been "face value" . This was followed by 10 definitions, such as "a legal way of reducing the amount of tax a person or company would normally pay:T____ a____" . This particular item should be answered "tax avoidance" . The distribution of questions types and domains is shown in Table 1.

Table 1 Distribution of question types in pre/post–tests

Question type	International Finance questions	Management Accounting questions	Total
Abbreviation	5	5	10
Gap–fill	2	7	9
Definition	5	5	10
Total	12	17	29

In the pre–test, half of the items belonged to the sub–domain of Management Accounting, the other half to International Finance. To select the test terms, I constructed corpora in the same way as the students did for the interventions (described below) and selected the most salient vocabulary items; these corpora were also used to generate the vocabulary lists and quizzes for the control groups. A very similar post–test, containing the same items as the pre–test, in a different order, was administered at the end of the semester. A small number of dummy questions were introduced into both tests, so that participants did not have to answer exactly the same set of questions on both occasions.

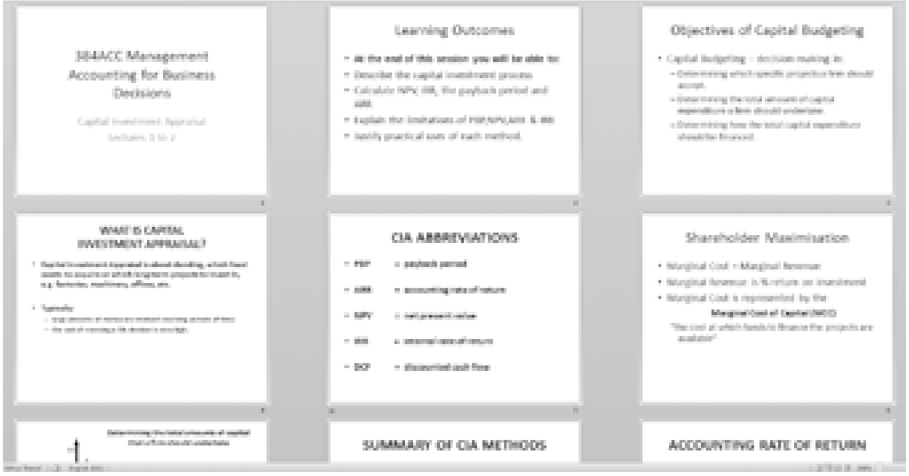
Learner Perception Questions

At the end of the interventions, and after the administration of the post-test, learners were asked to complete an anonymous online questionnaire about their experience in the class. They were asked for their view on the utility of the vocabulary learning methods to which they had been exposed, whether they would continue to use the resources they had created after the course, and whether they had acquired any skills other than language through the interventions. There were also some questions about the learning environment which are not immediately relevant to the present study. A qualitative analysis of the findings from the questionnaire is presented in the Results section.

The Interventions

Corpus Construction

In the first three of the 11 weekly classes, the experimental participants created and consulted their own corpora. The corpora were generated from lecture PowerPoints, seminar discussion notes, past test papers (sometimes with answers) and other materials provided by teachers in the AFIB department for students' use on the course Virtual Learning Environment (VLE, in this case Moodle) . Figure 2 shows a typical lecture PowerPoint, which includes learning outcomes, objectives, definitions and explanation of abbreviations, providing a rich set of domain keywords.



384ACC Management Accounting for Business Decisions

Capital Investment Appraisal

Lectures 1 to 2

Learning Outcomes

- At the end of this session you will be able to:
- Describe the capital investment process
- Calculate NPV, IRR, the payback period and ARR
- Explain the limitations of PBP, NPV, ARR & IRR
- Justify practical uses of each method

Objectives of Capital Budgeting

- Capital Budgeting – decision making in:
  - Determining which specific projects a firm should accept
  - Determining the total amount of capital expenditure a firm should undertake
  - Determining how the total capital expenditure should be financed

WHAT IS CAPITAL INVESTMENT APPRAISAL?

- [illegible]
- Typically:
  - [illegible]
  - [illegible]

CIA ABBREVIATIONS

- PBP = payback period
- ARR = accounting rate of return
- NPV = net present value
- IRR = internal rate of return
- DCF = discounted cash flow

Shareholder Maximisation

- Marginal Cost = Marginal Revenue
- Marginal Revenue is % return on investment
- Marginal Cost is represented by the Marginal Cost of Capital (MCC)

"The cost at which funds to finance the projects are available"

Determining the total amount of capital that a firm should undertake

SUMMARY OF CIA METHODS

ACCOUNTING RATE OF RETURN

Figure 2 Management Accounting lecture slides

As each new week' s lecture slides and seminar notes were made available on Moodle, the students would either add in the new content and grow their corpus, or create a new one.

The procedure for constructing a corpus (and consulting it) is shown in Figure 3. First, the user uploads the text content of teaching materials to form a mini-corpus, using the Sketch Engine: it is possible to upload files in a range of formats, including Word, PDF, zip and text, but PowerPoint files need to be converted to another format first. Because of the nature of lecture slides, the resulting corpus may not contain many full sentences, but it will include the key vocabulary for the particular topic. [[If this paper is accepted, a link to a YouTube explanatory screencast for students will be included here. The video is not anonymous]]

Students could opt to create a very specialized corpus, consisting of perhaps just one or two PowerPoints, for example on “Capital Investment Appraisal”（to which two lectures were devoted）. Alternatively they might decide to create a whole-module corpus, such as “Management Accounting”.

The Sketch Engine software is then used to generate a list of the most salient words—the keywords—in the corpus（words found frequently in the corpus, which are not found in a Sketch Engine defined reference corpus）. Thus, the word *the* is not salient, because it is found with equal normalized frequency in both specialist and reference corpora. The BootCat software（Baroni & Bernardini 2004; Baroni et al. 2006; available in Sketch Engine, or downloadable from http://bootcat.sslmit.unibo.it/）is then used to bootstrap a much larger corpus, consisting of texts from the web. Figure 3 illustrates this process.

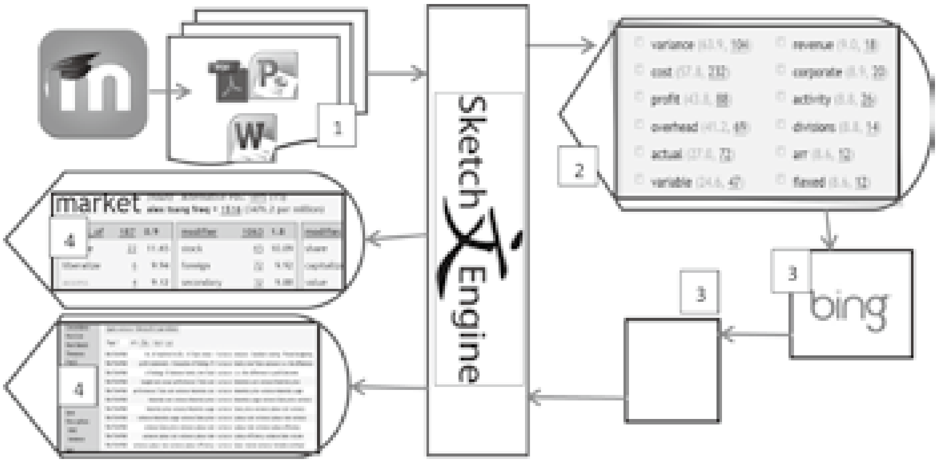
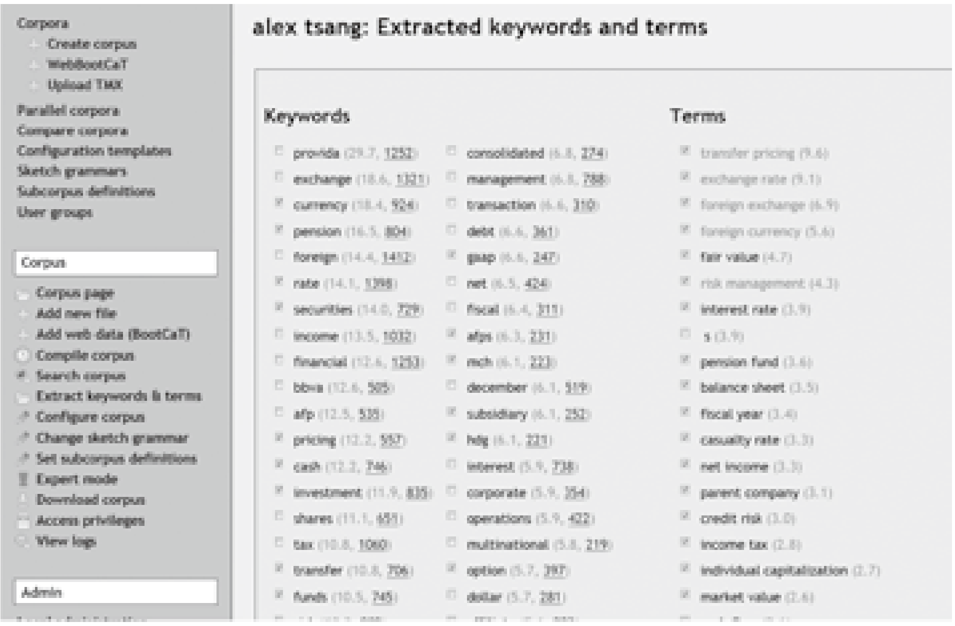


Figure 3 Schematic of corpus construction and consultation.

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

Construction of the bootstrapped, expanded corpus 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. This is one of the most crucial parts of the intervention, since it is here that the learners need to supply the keywords to seed（bootstrap） the expanded corpus. They do this by (1) inspecting the words and terms that Sketch Engine has determined to be salient in the original small corpus, (2) reflecting on whether they are in fact salient to the corpus domain, and (3) checking a box to show this is the case before submitting them to Sketch Engine as seed words. Figure 4 shows a student' s display at the point where he has made his selection and is about to submit.



alex tsang: Extracted keywords and terms

Keywords		Terms
☐ provida (29.7, 1252)	☐ consolidated (6.8, 274)	☒ transfer pricing (9.6)
☐ exchange (18.6, 1321)	☐ management (6.8, 788)	☒ exchange rate (9.1)
☒ currency (18.4, 924)	☐ transaction (6.6, 310)	☒ foreign exchange (6.9)
☒ pension (16.5, 804)	☐ debt (6.6, 361)	☒ foreign currency (5.6)
☐ foreign (14.4, 1452)	☒ gaap (6.6, 247)	☒ fair value (4.7)
☒ rate (14.1, 1398)	☐ net (6.5, 424)	☒ risk management (4.3)
☒ securities (14.0, 729)	☐ fiscal (6.4, 311)	☒ interest rate (3.9)
☐ income (13.5, 1032)	☒ afps (6.3, 231)	☐ s (3.9)
☐ financial (12.6, 1253)	☒ mch (6.1, 223)	☒ pension fund (3.6)
☐ bbva (12.6, 505)	☐ december (6.1, 519)	☒ balance sheet (3.5)
☐ afp (12.5, 535)	☒ subsidiary (6.1, 252)	☒ fiscal year (3.4)
☒ pricing (12.2, 557)	☒ hdg (6.1, 221)	☒ casualty rate (3.3)
☒ cash (12.2, 746)	☐ interest (5.9, 738)	☒ net income (3.3)
☒ investment (11.9, 835)	☐ corporate (5.9, 254)	☒ parent company (3.1)
☐ shares (11.1, 651)	☐ operations (5.9, 422)	☒ credit risk (3.0)
☐ tax (10.8, 1060)	☐ multinational (5.8, 219)	☒ income tax (2.8)
☒ transfer (10.8, 706)	☒ option (5.7, 397)	☒ individual capitalization (2.7)
☒ funds (10.5, 745)	☐ dollar (5.7, 281)	☒ market value (2.6)

Figure 4　Student' s Transfer pricing corpus at intermediate stage of construction

Note from Figure 4 that the MWUs are shown separately as *terms*. In this case, all are two-word terms, but longer MWUs do sometimes emerge as salient. The student has taken the opportunity to ignore spurious items such as *s*, showing an awareness of the unexpected, characterized by Charles (2012: 97) as "a key feature of corpus work" in construction of a DIY corpus. He has also eliminated words which he considers perhaps not specific enough, such as *foreign*. Note that a number of on-domain technical abbreviations have also been ticked, such as GAAP (Generally Accepted Accounting Principles) .

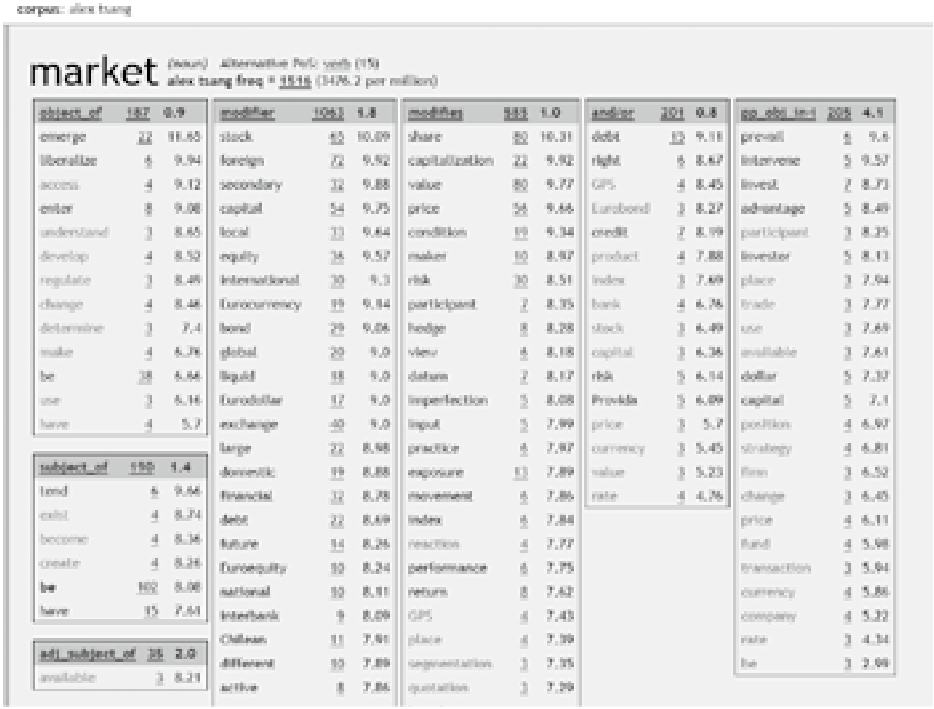
Corpus Consultation

The expanded corpus can be used in the following ways:

1. To produce lists of subject area words and terms for study.

2. To view Word Sketches, which give a one-page view of the collocations and grammatical structures in which a word or term participates. Figure 5 shows the Word Sketch for *market*; the reader will note that the most salient collocates are *emerge* in the (incorrectly assigned) object_of relation (*emerging market*) , *stock* as a modifier of *market* (*stock market*) , and *share* in the modified relation (*market share*) . Clicking on the underlined

frequency statistic yields a concordance for that particular collocation.



corpus: alex tsang

market (noun) Alternative PoS: verb (15)
alex tsang freq = 1516 (3476.2 per million)

object_of	187	0.9
emerge	22	11.65
liberalize	6	9.94
access	4	9.12
enter	8	9.08
understand	3	8.65
develop	4	8.52
regulate	3	8.49
change	4	8.46
determine	3	7.4
make	4	6.76
be	38	6.66
use	3	6.16
have	4	5.7

subject_of	150	1.4
tend	6	9.66
exist	4	8.74
become	4	8.36
create	4	8.26
be	102	8.08
have	15	7.64

adj_subject_of	35	2.0
available	3	8.21

modifier	1063	1.8
stock	65	10.09
foreign	72	9.92
secondary	32	9.88
capital	54	9.75
local	33	9.64
equity	36	9.57
international	30	9.3
Eurocurrency	19	9.14
bond	29	9.06
global	20	9.0
liquid	18	9.0
Eurodollar	17	9.0
exchange	40	9.0
large	22	8.98
domestic	19	8.88
financial	32	8.78
debt	22	8.69
future	14	8.26
Euroequity	10	8.24
national	10	8.11
interbank	9	8.09
Chilean	11	7.91
different	10	7.89
active	8	7.86

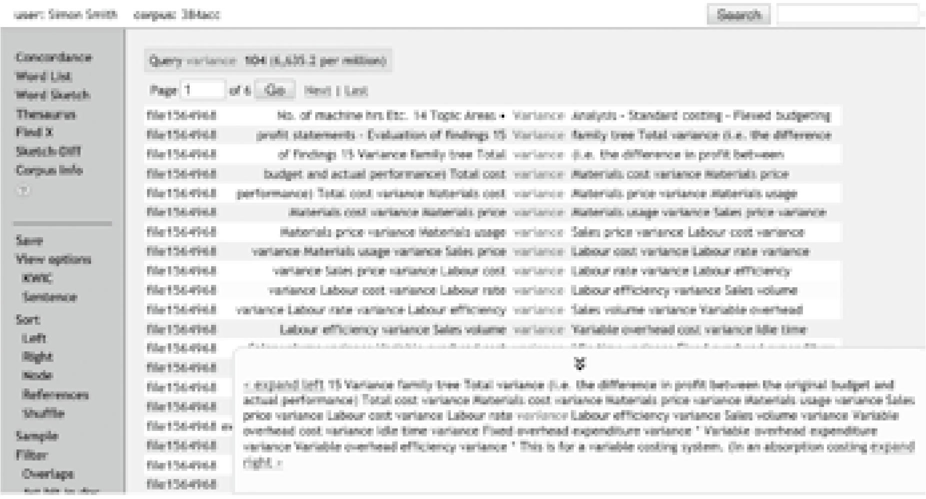
modifies	583	1.0
share	82	10.31
capitalization	22	9.92
value	80	9.77
price	56	9.66
condition	19	9.34
maker	10	8.97
risk	30	8.51
participant	7	8.35
hedge	8	8.28
view	6	8.18
datum	7	8.17
imperfection	5	8.08
input	5	7.99
practice	6	7.97
exposure	13	7.89
movement	6	7.86
index	6	7.84
reaction	4	7.77
performance	6	7.75
return	8	7.62
GPS	4	7.43
place	4	7.39
segmentation	3	7.35
quotation	3	7.29

and/or	201	0.8
debt	13	9.11
right	6	8.67
GPS	4	8.45
Eurobond	3	8.27
credit	7	8.19
product	4	7.88
index	3	7.69
bank	4	6.76
stock	3	6.49
capital	3	6.36
risk	5	6.14
Provida	5	6.09
price	3	5.7
currency	3	5.45
value	3	5.23
rate	4	4.76

pp_obj_in-i	205	4.1
prevail	6	9.6
intervene	5	9.57
invest	7	8.73
advantage	5	8.49
participant	3	8.25
investor	5	8.13
place	3	7.94
trade	3	7.77
use	3	7.69
available	3	7.61
dollar	5	7.37
capital	5	7.1
position	4	6.97
strategy	4	6.81
firm	3	6.52
change	3	6.45
price	4	6.11
fund	4	5.98
transaction	3	5.94
currency	4	5.86
company	4	5.22
rate	3	4.34
be	3	2.99

Figure 5 A Sketch Engine word sketch, showing principal grammatical relations of a keyword

3. To view a KWIC concordance of the words or terms focused on. Figure 6 shows how a given concordance line centring on a keyword may be selected for expansion.



Query variance 104 (6,635.2 per million)

file1564968	No. of machine hrs Etc. 14 Topic Areas •	Variance	Analysis - Standard costing - Flexed budgeting
file1564968	profit statements - Evaluation of findings 15	Variance	family tree Total variance (i.e. the difference
file1564968	of findings 15 Variance family tree Total	variance	(i.e. the difference in profit between
file1564968	budget and actual performance) Total cost	variance	Materials cost variance Materials price
file1564968	performance) Total cost variance Materials cost	variance	Materials price variance Materials usage
file1564968	Materials cost variance Materials price	variance	Materials usage variance Sales price variance
file1564968	Materials price variance Materials usage	variance	Sales price variance Labour cost variance
file1564968	variance Materials usage variance Sales price	variance	Labour cost variance Labour rate variance
file1564968	variance Sales price variance Labour cost	variance	Labour rate variance Labour efficiency
file1564968	variance Labour cost variance Labour rate	variance	Labour efficiency variance Sales volume
file1564968	variance Labour rate variance Labour efficiency	variance	Sales volume variance Variable overhead
file1564968	Labour efficiency variance Sales volume	variance	Variable overhead cost variance Idle time
file1564968			
file1564968			
file1564968			
file1564968			
file1564968			
file1564968			
file1564968			
file1564968			

‹ expand left 15 Variance family tree Total variance (i.e. the difference in profit between the original budget and actual performance) Total cost variance Materials cost variance Materials price variance Materials usage variance Sales price variance Labour cost variance Labour rate variance Labour efficiency variance Sales volume variance Variable overhead cost variance Idle time variance Fixed overhead expenditure variance * Variable overhead expenditure variance Variable overhead efficiency variance * This is for a variable costing system. (In an absorption costing expand right ›

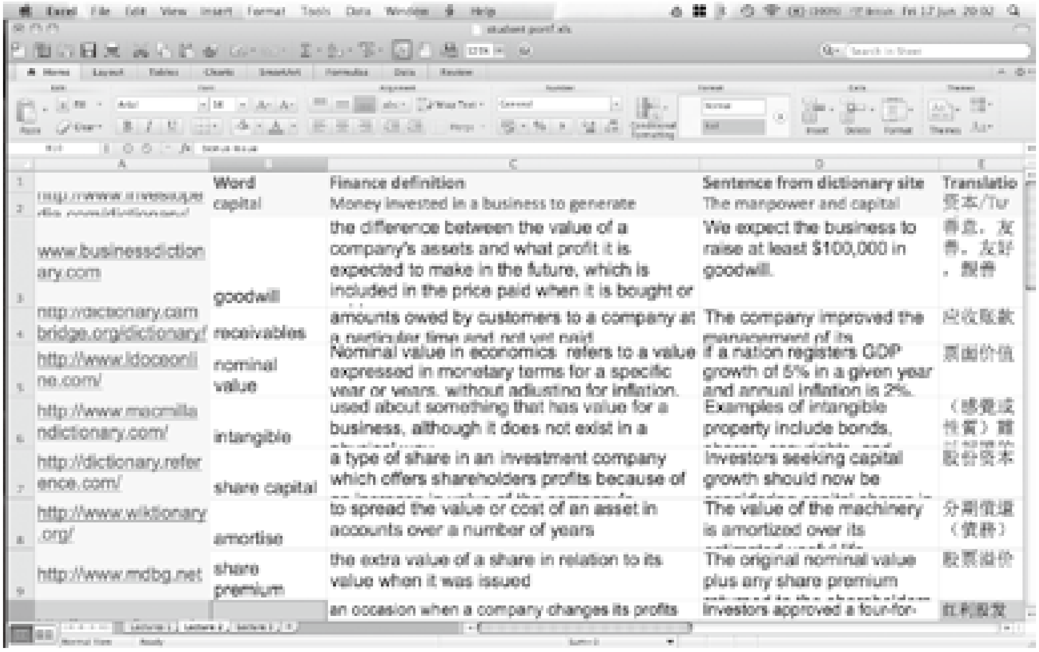
Figure 6 Sketch Engine concordance output

4. The student may also click on the fileNNNNNNN links seen in the concordance output

to refer back to the original texts of which the corpus is composed (from the expanded Web corpus, or from the student' s Moodle corpus) .

Vocabulary Portfolios

From the fourth week onwards, students were also asked to create and work with personal vocabulary portfolios. The portfolios took the form of an Excel spreadsheet, with a template issued by the instructor, as shown in Figure 7. In the figure, the top two rows (the header row, and the example *capital*) , and the leftmost column, consisting of links to dictionaries and other online resources, were supplied by the instructor. These areas of the spreadsheet are grey-shaded in Figure 7. The student has completed the second column with words and terms from their personal management accounting corpus, and the remaining columns with definitions and example sentences from online dictionaries, as well as translations into Chinese, the student' s L1. Students were also encouraged to take example sentences from their personal corpora, although this was less widely taken up.



	A	B	C	D	E
1		Word	Finance definition	Sentence from dictionary site	Translatio
2	http://www.investope dia.com/dictionary/	capital	Money invested in a business to generate	The manpower and capital	资本/Tu
3	www.businessdictionary.com	goodwill	the difference between the value of a company's assets and what profit it is expected to make in the future, which is included in the price paid when it is bought or	We expect the business to raise at least $100,000 in goodwill.	善意，友善，友好，股誉
4	http://dictionary.cambridge.org/dictionary/	receivables	amounts owed by customers to a company at a particular time and not yet paid	The company improved the management of its	应收账款
5	http://www.ldoceonline.com/	nominal value	Nominal value in economics refers to a value expressed in monetary terms for a specific year or years, without adjusting for inflation.	If a nation registers GDP growth of 5% in a given year and annual inflation is 2%.	票面价值
6	http://www.macmillandictionary.com/	intangible	used about something that has value for a business, although it does not exist in a	Examples of intangible property include bonds,	（感觉或性質）難以捉摸
7	http://dictionary.reference.com/	share capital	a type of share in an investment company which offers shareholders profits because of	Investors seeking capital growth should now be	股份资本
8	http://www.wiktionary.org/	amortise	to spread the value or cost of an asset in accounts over a number of years	The value of the machinery is amortized over its	分期偿還（债券）
9	http://www.mdbg.net	share premium	the extra value of a share in relation to its value when it was issued	The original nominal value plus any share premium	股票溢价
			an occasion when a company changes its profits	Investors approved a four-for-	红利股发

Figure 7　Student' s vocabulary portfolio excerpt

Control Group Tasks

Three PowerPoint vocabulary quizzes, with gapped KWIC concordances, were developed using domain vocabulary corpora created by the instructor. These were administered using PowerPoint in a similar way to the pre/post-tests, but scores were not recorded. Furthermore, a vocabulary list, generated by the instructor from a corpus containing each week' s ACC or

FIN (depending on the group) materials, was placed on Moodle weekly.

RESULTS & DISCUSSION

Of the 94 students enrolled for the modules, 55 were present for both the pre- and post-tests: 33 in the two control group classes, and 22 in the experimental group. This rather low response was because of attendance issues at the end of the semester. Lee & Swales (2006) encountered a similar problem in their corpus construction and analysis task, and were eventually led to abandon their planned post-test. I did not take this step, but low sample size is an unfortunate occasional corollary of classroom-based research. In the present study, the numerical improvements in performance that the approach suggests were mostly not found to be statistically significant.

Table 2 Results from pre- and post-tests

Group	Pre-test mean correct			Post-test mean correct			Improvement (**p = 000,*p < 0.05)		
	ACC	FIN	ALL	ACC	FIN	ALL	ACC	FIN	ALL
EFA1	4.04	1.70	5.74	6.97	4.00	10.97	2.93**	2.30*	5.23**
EFA2	3.80	1.54	5.34	5.73	3.29	9.02	1.93*	1.75*	3.68*
EFA3	5.18	2.12	7.29	7.03	4.12	11.15	1.85*	2.00*	3.86**
EFA4	4.59	2.50	7.09	6.53	3.97	10.50	1.94**	1.47*	3.41**
Experimental groups (EFA1 & 2)	3.91	1.61	5.52	6.29	3.61	9.91	2.38**	2.00**	4.39**
Control groups (EFA3 & 4)	4.89	2.30	7.19	6.79	4.05	10.83	1.9**	1.75**	3.64**
ACC groups (EFA1 & 3)	4.76	1.96	6.72	7.01	4.07	11.08	2.25**	2.11**	4.36**
FIN groups (EFA2 & 4)	4.25	2.09	6.34	6.18	3.68	9.86	1.93**	1.59**	3.52**
Overall	4.50	2.03	6.53	6.59	3.87	10.46	2.09**	1.84**	3.93**

Table 2 summarizes the results, giving mean scores in the pre-test and post-test, and the performance improvement (the difference between the pre- and post-test scores) . Mean scores are given for each class group (EFA1-4) , as well as the combined scores for experimental and control groups, and for the groups focusing on Management Accounting and on International Finance. The columns headed FIN and ACC represent mean scores on International Finance and Management Accounting questions respectively, and ALL is the sum of the two. It will be observed that the scores for FIN questions are lower than those for ACC questions across the board, despite the fact that the FIN scores were adjusted to reflect the lower number of questions; this is probably because the students had been exposed to many of the more general Management Accounting terms in earlier studies, while International Finance, especially with its focus on European markets, was a new field to them. Notwithstanding that, over the period of intervention one would expect roughly equal improvement in scores in both domains, and the improvement scores for FIN and ACC reflect that.

A t–test assuming unequal variance was conducted to measure the improvement on post– over pre–test. All four class groups (EFA1–4) registered post–test scores which were a significant improvement on pre–test scores, as shown in Table 2. Thus, H1 (All groups will perform better in the post–test than they did in the pre–test) was supported.

An unexpected finding was that experimental groups performed less well than control groups on both pre– and post–tests, with the former registering a greater improvement on the post–test. This appears to indicate that the experimental groups consisted of slightly weaker students, and since the groups were assigned arbitrarily there is no immediate explanation. However, these differences were not statistically significant.

A 2–way univariate ANOVA was run in SPSS to compare the pre–post test improvement of various groups, but no differences were found to be significant. Experimental groups saw greater improvement than control groups, so there is some limited support for H2. H3 (that both ACC groups would register greater improvement on ACC items than FIN groups) was partially supported: EFA1 outperformed EFA2, but EFA3 improved less than EFA4. H4 was not supported, as EFA1 and EFA3 both outperformed EFA2 and EFA4 respectively.

For improvement in ACC scores, there was no difference between experiment and control groups for students in the FIN domain. There was a greater improvement for ACC domain students in the experiment group compared to ACC domain students in the control group (improvement of 2.93 compared to 1.85) . For improvement in FIN scores, students in both domains saw a greater improvement if they were in the experiment group compared to control (improvement of 2.3 compared to 2.0 for ACC domain; 1.75 compared to 1.47 for FIN domain) . Thus, ACC domain students did improve more in ACC scores than they did in FIN scores, and FIN domain students did improve more in FIN scores than they did in ACC scores. Again, the differences were not statistically significant.

Effect sizes were calculated for the improvement scores of the control and experimental groups, and are shown in Table 3. The figures represent the improvement scores shown in Table 2 divided by the standard deviation in the improvement score for the control group (equal to 1.48 for ACC questions, 1.72 for FIN questions, and 2.24 overall) . This way of calculating effect size is described by Coe（2002）.

Table 3 Effect sizes for improvement scores of control and experimental groups

	ACC	FIN	ALL
Experimental groups (EFA1 & 2)	1.60	1.16	1.96
Control groups (EFA3 & 4)	1.20	1.01	1.62

The table shows that the effect sizes due to the corpus treatment are indeed larger than those found in the control groups, which used the vocabulary list treatment. For ACC questions to the experimental group, the effect size of 1.6 means that there would be a 79% chance of correctly assigning a random experimental group student to the experimental group, rather than to the control group; that 95% of the control group would have a lower improvement score than the average experimental group score; and that given a control group of 25, only the highest scoring control group member would be likely to attain the average experimental group score. These interpretations are taken from Coe（2002）.

Results from perceptions questionnaire

In all, 60 participants out of 94 completed the post-intervention perceptions questionnaire—34 from the experimental groups and 26 from the control groups. Because the control and experimental groups had been subjected to different teaching approaches, a direct comparison of the success of the methods was not possible. Both experimental and control group respondents appeared to be satisfied with the different aspects of their interventions, as Table 4 illustrates. There are no significant differences between the responses.

Table 4 Level of participant satisfaction with interventions

	Intervention	"Very useful" /" quite useful" responses	
		Number	%
Experimental groups	Making a corpus using files from Moodle	28	82.4
	Making a vocabulary portfolio	26	76.5
Control groups	Vocabulary lists	23	84.6
	Quizzes	23	84.6

When asked whether they would continue to use the corpus and vocabulary portfolio resources after the end of the course, 20 respondents（58.8%）claimed that they would, while 6（17.6%）stated that they would like to but were not sure how. All learners felt that they had acquired non-language/transferable skills, with 28 learners (70%) citing programs such as Office and Sketch Engine as the most significant, and the other 12 (30%) stating that their Web search skills had improved.

For the experimental group students, it seems to be in the construction of the corpus and the compilation of the vocabulary portfolio that the benefit of the approach lies. The Sketch Engine platform offers a number of tools for inspecting the data, and the students found the Word Sketch function of particular interest. Anecdotally, they appeared to have less patience to deal with concordance output, and it was somewhat difficult to convince them of the pedagogical value of examining sentence fragments; the reservations of Kilgarriff (2008) , Lamy & Klarskov Mortensen (2007) , and Yoon and Hirvela (2004) , noted in the Literature Review, probably apply to the participants of this study as well.

LIMITATIONS

As has been noted already, the limited sample size (occasioned partly by the relatively low number of participants in the post-test) and consequent lack of statistically significant improvement figures means that the results from the study, while promising, must be seen as tentative.

There is, of course, rather more to "knowing" a word or term than being able to match it to a definition or use it in a gap-fill, so it is not known how valid the pre- and post-test were as a means of measuring vocabulary knowledge: they might not capture the extent to which students benefited from the intervention, and it may be that the benefits of this type of activity are not readily tangible or measurable by tests.

A limitation of the questionnaire feedback should be noted: even though responses were anonymous, it is still likely that some of the students would have responded positively because they felt it was the right thing to do.

There were also some logistical difficulties with the approach. For one, it was quite difficult to schedule the vocabulary tasks with a crowded EAP syllabus to cover, in a class which only met for two hours per week. This meant that I was not able to spend as much time on the interventions as I would have liked. The process of logging in and finding the corpora or files that were being worked on the previous week tended to eat into class time, and students needed constant monitoring to keep them on task and away from online distractions.

Some tasks were assigned for self-study in the early part of the semester, with students being asked to feed back to the class by way of Moodle forums. With this particular cohort of students, however, other coursework commitments (and perhaps motivational factors) meant that students were happier working on tasks in the classroom. The presence of a facilitating teacher probably also provided reassurance for them.

CONCLUSION AND FUTURE PLANS

This paper has presented a DIY corpus construction and vocabulary portfolio compilation intervention for university students of academic English who major in Accounting & Finance. Improvements in technical vocabulary knowledge, it is tentatively concluded, were made as a result of the intervention, and students reported that the experience was beneficial. Encouraging effect sizes of the intervention were also found.

Improvements were mostly not, however, found to be statistically significant. For this, larger numbers are needed; fortunately, our inter-departmental collaborations mean that in future cohorts, it will be possible to repeat the experiment on a larger scale. EAP students in other disciplines (for example International Business, and Engineering) will be encouraged to create corpora and vocabulary portfolios as part of their Academic English modules. In our university context, this will bring the added benefit of extending the demographics of the study: virtually all AFIB students at our institution are from China, and it would be interesting to see whether the findings can be generalized to other cultural/L2 backgrounds (since other cohorts are more mixed in terms of nationalities) .

Although the organizational and logistical aspects were not formally evaluated, I believe that the challenges met with by the teacher were to some extent offset by benefits to the students, who had previously had very little experience of file management or any kind of professional/academic use of computer resources (other than word processing and web searches) . Jackson (1997) lists a number of such skills that his Computer Aided Text Analysis students acquired: project management, problem solving and report writing, as well as computer skills. The questionnaire results of the present study, indeed, suggest that useful transferable skills were acquired, and this is something that ought to be further explored in a future study.

Traditionally, quite a lot of CALL provision intended for lab use has consisted of gap-fill or drop-down menu tasks. Students tend to find these quite fun, but they may be of greater utility for mastering (say) the mechanics of paraphrasing, or the niceties of a particular tense, than discovery of relevant, domain-specific, authentic vocabulary in context. The corpus-informed lexical resource creation tasks described here provide a motivating and meaningful way for students to access and learn the terminology and usages of their own specialist subjects.

REFERENCES:

[1] Ädel, A. (2010). Using corpora to teach academic writing: Challenges for the direct approach. In M. C. Campoy-Cubillo, B. Belles-Fortuno, & M. L. Gea-Valor (Eds.), *Corpus-based approaches to ELT* (pp.

39–55). London: Continuum.

[2] Aston, G. (2002). The learner as corpus designer. In B. Kettemann, & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 9–25). Amsterdam: Rodopi. Retrieved August 27, 2009, from http://www.sslmit.unibo.it/~guy/graz.htm.

[3] Author (2011).

[4] Author (2015).

[5] Baroni, M. & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the Web. In *Proceedings of 4th*InternationalConferenceon LanguageResourcesand.

[6] *Evaluation*. Lisbon, Portugal, 1313–1316. Retrieved August 27, 2009, from http://people.sslmit.unibo.it/~baroni/publications/lrec2004/bootcat_lrec_2004.pdf.

[7] Baroni, M., Kilgarriff, A., Pomik á lek, J., Rychlý, P. (2006) WebBootCaT: instant domain-specific corpora to support human translators. *Proceedings, 11th Annual Conference of the European Association for Machine Translation Conference*. Oslo, Norway, 247–252. Retrieved August 27, 2009, from http://www.mt-archive.info/EAMT-2006-Baroni.pdf.

[8] Bernardini, S. (2000). Systematising serendipity: Proposals for concordancing large corpora with language learners. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 225–234). Frankfurt: Peter Lang.

[9] Boulton, A. (2009). Testing the limits of data-driven learning: Language proficiency and training. *ReCALL*,*21*(1), 37–54.

[10] Boulton, A. (2008). Bringing corpora to the masses: free and easy tools for language learning. In N. K ü bler (Ed.), *Corpora, Language, Teaching, and Resources: From Theory to Practice* Bern: Peter Lang. Retrieved July 7, 2016, from http://hal.archives-ouvertes.fr/docs/00/32/69/80/PDF/XXXX_boulton_TaLC_interdisciplinary.pdf.

[11] Browne, C., Culligan, B., & Phillips, J. (2013). *A New Academic Word List*. Retrieved July 11, 2016, from http://www.newacademicwordlist.org/.

[12] Campion, M. & Elley, W. (1971) *An academic vocabulary list*. Wellington: New Zealand Council for Educational Research.

[13] Castagnoli, S. (2006). Using the Web as a source of LSP corpora in the terminology classroom”. In M. Baroni & S. Bernardini (Eds.), *Wacky! Working papers on the Web as corpus* (pp. 159–172). Bologna: Gedit. Retrieved July 11, 2016, from http://wackybook.sslmit.unibo.it/pdfs/castagnoli.pdf.

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

[15] Charles, M. (2014). Getting the corpus habit: EAP students’ long-term use of personal corpora.*English for Specific Purposes*,*35*, 30–40.

[16] Coe, R. (2002). It’s the Effect Size, Stupid: What effect size is and why it is important. Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, 12–14

September. Retrieved July 14, 2016, from http://www.leeds.ac.uk/educol/documents/00002182.htm.

[17] College Entrance Examination Center. (2002). 大学入学考试中心高中英文参考词汇表 [High School English Reference Wordlist]. Retrieved July 12, 2016, from http://www.ceec.edu.tw/Research/paper_doc/ce37/2.pdf.

[18] Coxhead, A. (2000). A new academic word list.*TESOL Quarterly*,*34*(2), 213–238.

[19] Gardner, D., & Davies, M. (2013). A new academic vocabulary list.*Applied Linguistics*, 35: 1–24.

[20] Gavioli, L. (2009). Corpus analysis and the achievement of learner autonomy in interaction. In L. Lombardo (Ed.), *Using corpora to learn about language and discourse* (pp. 39–71). Bern, Switzerland: Peter Lang.

[21] Ghadessy, P. (1979). Frequency counts, words lists, and materials preparation: A new approach. *English Teaching Forum 17,* 24–27.

[22] Gilner, L. (2011). A primer on the General Service List.*Reading in a Foreign Language*,*23*(1), 65–83.

[23] Hirsh, D., & Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, *8*, 689–696.

[24] Hyland, K. and Tse, P. (2007). Is there an “Academic Vocabulary”? *TESOL Quarterly*, *41*(2): 235–253.

[25] Jackson, H. (1997). Corpus and concordance: Finding out about style. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.) *Teaching and language corpora* (pp. 224–239), London: Longman.

[26] Johns, T. (1986). Micro–Concord: a language learner’s research tool. *System*, *14*(2): 151–162.

[27] Johns, T. (1991). Should you be persuaded: Two examples of data–driven learning. In Johns, T.F. and King, P. (Eds.) *Classroom concordancing* (pp. 1–13), Birmingham: ELR.

[28] Kilgarriff, A., Husak, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the 11th EURALEX International Congress,* Barcelona, Catalonia. Retrieved August 28, 2009, from http://www.kilgarriff.co.uk/Publications/2008–KilgEtAl–euralex–gdex.doc.

[29] Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress,* Lorient, France. Retrieved August 28, 2009, from http://kilgarriff.co.uk/Publications/2004–KilgRychlySmrzTugwell–SkEEuralex.rtf.

[30] Lamy M–N. & Klarskov Mortensen H. J. (2012). Using concordance programs in the Modern Foreign Languages classroom. Module 2.4 in Davies G. (ed.) *Information and Communications Technology for Language Teachers (ICT4LT)*, Slough, Thames Valley University [Online]. Available at: http://www.ict4lt.org/en/en_mod2–4.htm [Accessed 11 July 2016].

[31] 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.

[32] Lynn, R. W. (1973). Preparing word lists: a suggested method. *RELC Journal 4*(1), 25–32.

[33] Martinez, R., & Schmitt, N. (2012). A phrasal expressions list.*Applied Linguistics*, *33*(3), 299–320.

[34] Mudraya, O. (2006) Engineering English: A lexical frequency instructional model. *English for Specific Purposes, 25*(2), 235–256.

[35] Neufeld, S. (2012). *Data driven language learning: Using corpus linguistics in language learning.* Retrieved July 11, 2016, from http://www.scoop.it/t/data–driven–language–learning.

[36] Praninskas, J. (1972) *American university word list.* London: Longman.

[37] Regents of the University of Michigan (2011). *MICASE Kibbitzers*. Retrieved July 11, 2016, from https://web.archive.org/web/20111008033810/http://micase.elicorpora.info/micase–kibbitzers.

[38] Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge: Cambridge University Press.

[39] Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know?*TESOL Quarterly*, 145–171.

[40] Simpson–Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research.*Applied Linguistics*,*31*(4), 487–512.

[41] Stubbs, M. (2002). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

[42] Thurstun, J., & Candlin, C. N. (1997).*Exploring Academic English: A Workbook for Student Essay Writing*. Sydney: Macquarie University.

[43] Tribble, C. and Jones, G. (1990) *Concordances in the Classroom: A Resource Book for Teachers*. Harlow: Longman.

[44] Tyne, H. (2009). Corpus oraux par et pour l' apprenant [Spoken corpora by and for the learner]. In A. Boulton (Ed.), *Des documents authentiques oraux aux corpus: Questions d'apprentissage en didactique des langues* (pp. 91–111). Nancy, France: M é langes CRAPEL. Retrieved July 12, 2016, from https://hal.archives–ouvertes.fr/hal–00416544/document.

[45] Van den Branden, K. (2013). Task–based language teaching and learning, In P. Robinson (Ed.), *The Routledge encyclopedia of second language acquisition* (pp. 628–632). New York: Routledge.

[46] Vincent, B. (2013). Investigating academic phraseology through combinations of very frequent words: a methodological exploration.*Journal of English for Academic Purposes12*(1), 44–56.

[47] Wang, J., Liang, S. & Ge, G. (2008) Establishment of a Medical Academic Word List. *English for Specific Purposes 27*(4), 442–458.

[48] West, M. (1953). *A general service list of English words*. London: Longman.

[49] Xue, G., & Nation, P. (1984). A university word list.*Language learning and communication 3*(2), 215–229.

[50] Yoon, H. and Hirvela, A. (2004) ESL student attitudes toward corpus use in L2. *Journal of second language writing*, *13*(4): 257–283.

[51] Zanettin, F. (2002) DIY Corpora: The WWW and the Translator. In Maia, B., Haller, J., & Urlrych, M. (eds.) *Training the Language Services Provider for the New Millennium*. Porto: Facultade de Letras, Universidade do Porto, 239–248.396396.

作者简介：

Simon Smith, PhD, FHEA, Assistant Professor School of Humanities and Research Associate Centre for Global Learning: Education and Attainment. He is a member of two research groups: Corpus Linguistics at Coventry（CLaC）and Centre for Global Learning: Education and Attainment（GLEA）. His research is on the use of corpora for language teaching and learning. Recent published work has focused on both Chinese and English learning. Research interests include Corpus query tools, Computational linguistics, Chinese teaching and learning, English for Academic Purposes（EAP）, Data-driven language learning（DDL）, Corpora in language teaching and learning, and Computer assisted language learning（CALL）. His recent publications include 7 papers in the authoritative periodicals of international significance.