DIY Corpora for Accounting & Finance Vocabulary Learning

Smith, S

Original citation & hyperlink:
Smith, S 2020, 'DIY Corpora for Accounting & Finance Vocabulary Learning', English for Specific Purposes, vol. 57, pp. 1-12
https://dx.doi.org/10.1016/j.esp.2019.08.002

DOI 10.1016/j.esp.2019.08.002
ISSN 0889-4906

Publisher: Elsevier

DIY Corpora for Accounting & Finance Vocabulary Learning

ABSTRACT

It has been shown that language learners can benefit from a discovery-based learning process whereby they construct as well as consult their own specialist corpora and vocabulary portfolios, for the purposes of translator training (Castagnoli 2006), for general English (Author 2011) and for academic English learning (Charles 2012, Author 2015a).

In the present study, a cohort of 94 international students on an EAP module, majoring in Accounting and Finance, was divided into hands-on (treatment) and hands-off (control) groups. Both groups were subjected to a pre-test consisting of specialist terms that would be encountered on their course (not only in the EAP class, but also on the Accounting and Finance modules). The hands-on group spent about 20 minutes per weekly class constructing domain-specific DIY corpora and generate subject vocabulary portfolios. The results of a post-test indicated that the hands-on group had achieved a slightly greater improvement in domain vocabulary knowledge than the hands-off group (which used corpora and vocabulary lists provided by the teacher). A participant questionnaire showed that the students found the approaches useful for vocabulary learning.

Keywords: DIY corpus; data-driven learning; terminology; vocabulary; accounting; finance

1. 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 particular 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 use the Sketch Engine corpus analysis tool (Kilgarriff et al 2004;

www.sketchengine.eu) for this purpose. They then expand the corpus, using software tools provided, to add related texts from the Web. Next, they generate a list of the salient (as calculated by the Sketch Engine) 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, collocating words, 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).

2. LITERATURE REVIEW

2.1 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). The approach attempts to impart linguistic knowledge by making available samples of authentic language, from corpora, and inviting language learners to discover usage patterns for themselves. 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".

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 (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, considerably less glowingly, as *incidentalism* (whilst admitting to having adopted it in their own study). Whilst supporting the approach, Ädel (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 feelings of overload, given the unlimited range of potential actions that can be taken in DDL, "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 (2009: 44) finds 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". As has been noted, the students in the present study were tasked with constructing their own corpora, and developing wordlists based on them. I will return to the literature on corpus construction by learners in the final subsection of this literature review. Next, however, I look at corpus-informed wordlists and the surrounding literature.

2.2 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. Wang and Nation (2004) concur; they showed that the various senses of the words *volume* and *credit* are distributed differently across AWL disciplines.

Hyland & Tse (2007) investigated the distribution of AWL words in their own academic corpus, and found considerable variation in the ways words in different parts of speech 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 Academic Collocations List (Ackermann & Chen, 2013). 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.

2.3 Construction of Corpora by Learners

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) can motivate learners by conferring on them a feeling of ownership of the resource, 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].

Corpus compilation may help learners to acquire transferable skills as well as language. Boulton (2008) and Jackson (1997), for example, report that their students gained problem-solving and ICT (Information and communications technology) competencies. Charles (2014) found that some students were motivated enough to consult and even add to their corpus after the end of the course. Lee & Swales (2006) report that some of their students even purchased their own copies of Wordsmith Tools (Scott 2018), indicating a commitment to continuing with corpus construction and analysis in the future.

Castagnoli (2006) asked students of translation to use the BootCaT software (Baroni & Bernardini, 2004) to create corpora devoted to particular domains of their choice, and then to extract from them lists of keywords, used to generate terminology databases and glossaries. The more specialized the domain, she found, the more relevant terms it was possible to extract. Castagnoli gave students a technical translation task as an assessment, and instructed them to prepare for the task by building relevant web corpora, using BootCaT, and extracting a glossary of terms from it.

Author (2011) extended Castagnoli's approach to non-specialist language learners in a Taiwanese university. The construction of the corpus was first bootstrapped (or seeded) from keywords supplied by the user. The search engine API (application programming interface) module was used to find web pages that appear to address keyword-related topics; then other BootCat components extracted text from the web pages which was gathered into a corpus. Students were told to create and consult a corpus related to their own major subject, and give some analysis and commentary. One student commented as follows:

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.

Having given some account of the background to DDL, prior work on corpus-based wordlists, and the work on corpus construction by learners that particularly motivated the present study, let us turn now to the study itself.

3. METHODOLOGY

3.1 Preliminary Study

A group of six Accounting and Finance for International Business (AFIB) students constructed DIY corpora on an EAP (English for Academic Purposes) module, taught at a public university in the UK. The study is reported in greater detail by Author (2015a). As will be described in Section 3.3.1, participants built and consulted their own corpora, based on texts and learning materials that had been made available by their AFIB module tutors. They studied concordances and consulted

Word Sketches (one-page summaries of word usage) in the Sketch Engine corpus analysis tool, focusing on academic and accounting/ finance words and terms from their corpora.

Despite satisfaction with the approach reported by the students, observation of the students at work suggested that they 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 Ädel (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.

3.2 Main Study

The present quantitative study follows up Author (2015a), although on a considerably larger scale. Some further aspects of the methods adopted here were outlined in Author (2015b). The study ran over a period of one (11-week) second semester. The following research questions were 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.

3.2.1 Participants

A cohort of 94 Accounting and Finance for International Business students took part in the study. They were students at the same UK university as the participants in the preliminary study. Of the 94, 88 were native speakers of Chinese (Mandarin or Cantonese), and as with the preliminary study, they were IELTS level 6.5-7 users of English. There were altogether four class groups, of which two were designated *hands-on DDL* groups (47 students), while the other two classes acted as

hands-off DDL groups (also 47 students). The author taught the hands-on groups, in computer labs, while a colleague taught the hands-off groups in a traditional classroom.

Participants were asked to study two areas of specialist accounting and finance vocabulary. These areas corresponded to two of the home department modules studied by all participants. HOn1 (hands-on group 1) and HOff1 (hands-off group 1) explored Management Accounting (ACC) vocabulary; HOn2 and HOff2 focused on the domain of International Finance (FIN). ACC and FIN are two of the three content modules followed by all AFIB students in the second semester.

The hands-on groups built their own corpora, and used them to generate lists of Accounting & Finance terms, which were in turn incorporated into student vocabulary portfolios. This work was done in the computer lab, for around 20 minutes each week, guided by the author. The hands-off groups were each week handed lists of specialist Accounting & Finance vocabulary, generated from corpora which had been created by the author in the same way as the students in the hands-on groups. The hands-off students then used the lists to develop vocabulary portfolios, in the same way as the hands-on groups, but in self-study mode. These procedures are described further in the Interventions section below.

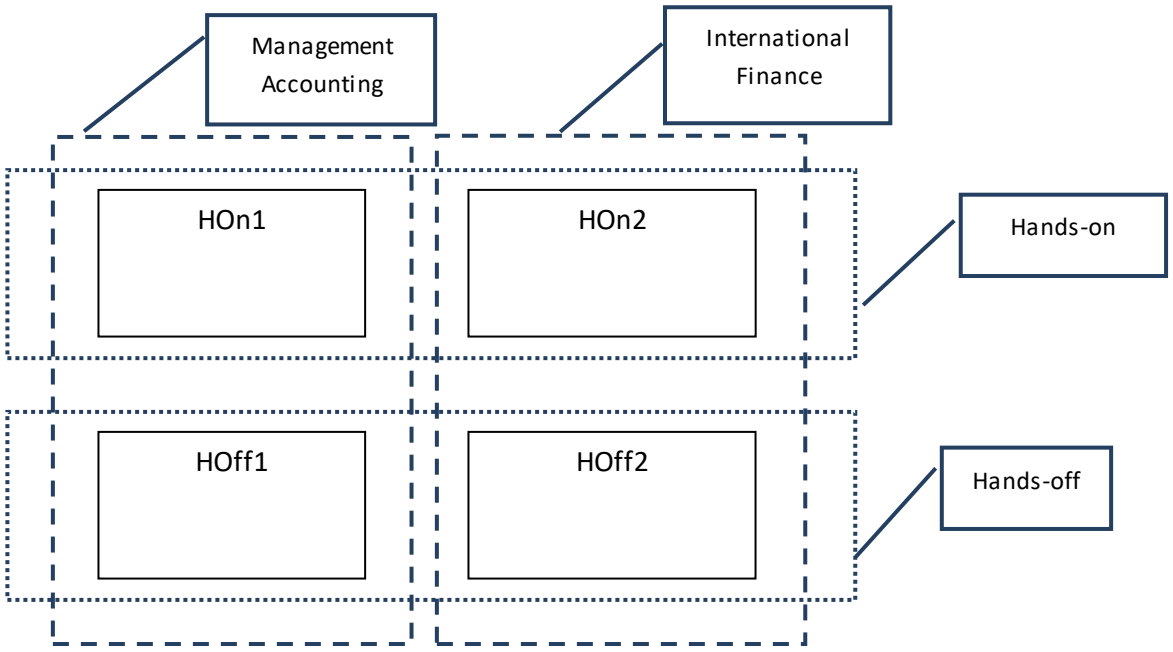

Figure 1 Arrangement of groups of participants

Figure 1 shows how the groups of participants were configured. It was predicted that:

H1. The performances of all groups in the post-test will surpass those of the pre-test.

H2. Hands-on DDL groups will show a greater improvement in the post-test than Hands-off groups.

H3. HOn1 will improve more than HOn2, and HOff1 more than HOff2, on Management Accounting vocabulary.

H4. HOn2 will improve more than HOn1, and HOff2 more than HOff1, on International Finance vocabulary.

3.2.2 Pre- and Post-tests

At the start of the semester, we administered a pre-test. This was intended to assess participants' prior knowledge of International Finance and Management Accounting vocabulary. The test featured acronyms (such as *SDR*), in which the students were asked what the initial letters stood

for (here, *Special Drawing Rights*). These were followed by gap-fills, for example "Some corporations use t_____ p_____ [transfer pricing] to make their profits seem lower than they really are".

There were also definition test items, such as "an agreement for the sale of currencies, goods, etc. at a fixed price to be given to a buyer on a future date: F_____ c_____" [Answer: Forward contract]. The distribution of questions types and domains is shown in Table 1. At first sight, the definitions may appear to the reader to be in a very similar format to the gap-fills; in the definitions, however, there is no information gap, with the semantic content of the answer being fully contained in the left-hand part of the item.

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

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

To select the test terms, the author 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 used by the hands-off DDL groups. The text of definitions and gap-fills was taken from Oxford Reference (2016), or in a few cases from the generated corpora. At the end of the intervention, a post-test was administered. It was similar to the pre-test, but with item presentation re-sequenced.

It was intended that there would be equal numbers of International Finance and Management Accounting questions. After the tests had been administered, however, it was discovered that two of the International Finance gap-fills in fact belonged to the domain of Management Accounting, and a further International Finance gap-fill was invalid. This is why the numbers of gap-fill items are unequal, as shown in Table 1.

3.2.3 Learner Perception Questions

At the end of the interventions, and after the administration of the post-test, learners were given an anonymous online questionnaire to complete, with information 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.

3.3 The Interventions

3.3.1 Corpus Construction

The participants' corpora were generated from lecturers' PowerPoint slides, seminar discussion notes, past test papers (sometimes with answers) and other materials provided for students' use by participants' home department instructors on the course Virtual Learning Environment (VLE, in this case Moodle). PowerPoint slideshows typically provide a rich set of domain keywords, including learning outcomes, objectives, definitions and explanation of abbreviations or acronyms.

As Accounting and Finance tutors added new content to the course Moodle week by week, participants could either build it in to expand their corpora, or create new corpora, depending on how specialized the individual student wished their resource to be.

Figure 2 shows the steps involved in constructing and then consulting a corpus. Firstly, the Sketch Engine is used to upload files consisting of course materials to create a mini-corpus: 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. [[If this paper is accepted, a link to a YouTube explanatory screencast for students will be included here. The video is not anonymous]]

Once the mini-corpus has been created, a list of the top-ranking (by default, 100) keywords and key multi-word units is generated from it. MWU status is determined by calculating the association scores of words in a collocation, using the algorithm logDice (Rychlý, 2008). After that, the Sketch Engine divides the normalized frequency of the words and MWUs in the corpus by the frequency of the same item in a reference corpus (enTenTen, a ten billion-word web corpus also available on Sketch Engine). This yields a keyness score for each word and MWU which determines the rank order in the list.

A tool known as BootCat (Baroni & Bernardini 2004; Baroni et al. 2006; available in Sketch Engine, or downloadable from http://bootcat.sslmit.unibo.it/) is then run, using the mini-corpus keywords to seed a further corpus, probably considerably larger, constructed from Web documents.

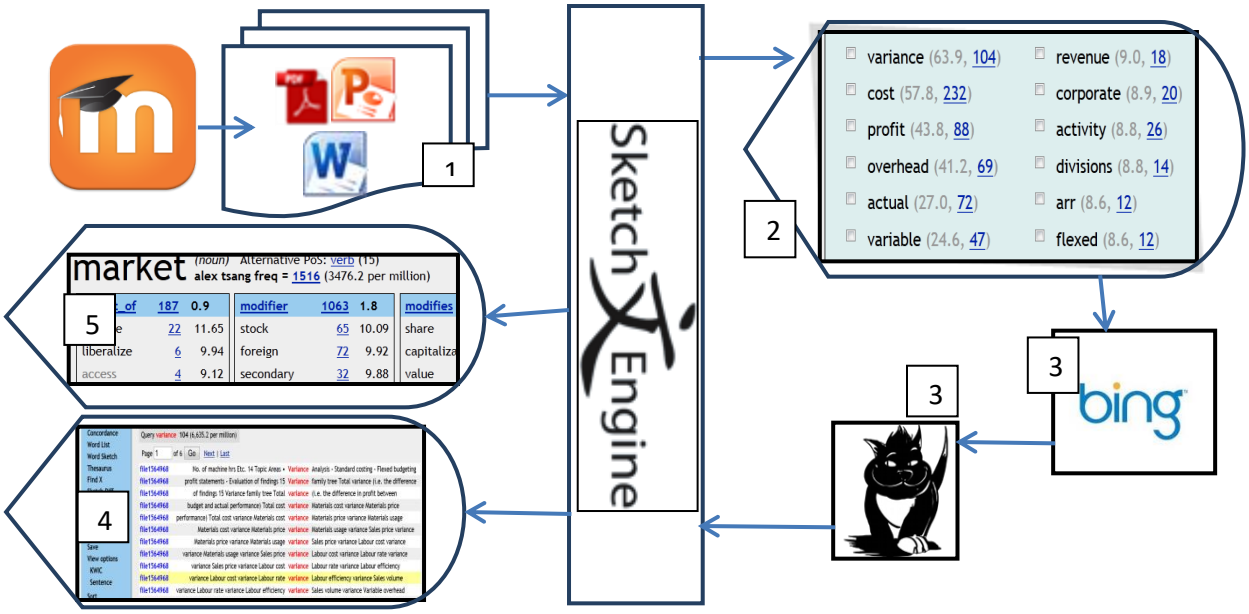

Figure 2 Flowchart showing construction of student DIY corpus, followed by consultation.

In Figure 2, [1] shows files being input to Sketch Engine to generate a wordlist [2]. This data is passed to Bing via the BootCat API [3] to create a further Sketch Engine corpus. Outputs such as concordances [4] and word sketch [5] may then be consulted. [This description should be kept with Figure 2]

Construction of the bootstrapped, expanded corpus is one of the most crucial parts of the intervention, since it is here that the hands-on 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 intuitively 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 3 shows a student's display at the point where he has made his selection and is about to submit.

Figure 3 Keywords and terms from student's *Transfer pricing* corpus at intermediate stage of construction. Checked boxes indicate student selected term for seeding expanded corpus. Underlined whole numbers indicate the frequency in the corpus, and if clicked on will link to a KWIC concordance of the word in question; the numbers to one decimal place indicate the keyness, as explained above, of the item.

When building the expanded corpus, there are a large number of optional parameters which can determine, among other things, the size of the corpus. Students were encouraged to use the default settings, for simplicity's sake, and the parameters are not discussed here. A student's original corpus would typically contain 20-40 thousand tokens, and an expanded corpus would often be around 10-15 times that size.

Note from Figure 3 that the MWUs are displayed by Sketch Engine as *terms*, while single words are labelled *keywords*. In this case, all of the terms consist of two words, but longer MWUs do sometimes emerge as salient. The student has taken the opportunity to ignore spurious items such as *provida* and *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. The student has also eliminated words which he considers perhaps not domain-specific enough, such as *foreign*. Note that a number of on-domain technical acronyms have also been ticked, such as GAAP (Generally Accepted Accounting Principles).

On the other hand, the student does appear to have missed some important words, salient to the domain, such as *multinational* and *transaction*. This is probably because the student did not know these words – it is a feature, or perhaps a limitation, of DDL that students are likely to learn more about the usage of words with which they have some familiarity than to acquire entirely new vocabulary. A more engaged student, perhaps, would notice, look up and ultimately acquire vocabulary identified as being salient to the domain by the Sketch Engine.

3.3.2 Corpus Consultation

The expanded corpus can be used to produce:

1. A further list of domain-specific keywords and terms (that is, words and MWUs).
2. Word Sketches, which provide a tabulated summary of the collocational and grammatical patterns of a word, as shown in Figure 4.

Figure 4 Sketch Engine Word Sketch output. Stock occurs as a modifier of market 65 times in the corpus, and is around 11.8 times more common in this student corpus than in the reference corpus (normalized for frequency).

3. KWIC concordances of the keywords or terms focused on. Figure 5 shows how a given concordance line centring on a keyword may be selected for expansion.

Figure 5 Sketch Engine concordance output

The student may also click on the link to the left of each KWIC line in the concordance output (as Figure 5), to link back to the original texts of which the corpus is composed (texts from the expanded Web corpus, or from the student's Moodle corpus).

3.3.3 Vocabulary Portfolios

From the fourth week of the intervention, the hands-on and hands-off participants were all asked to create and use personal vocabulary portfolios. The portfolios took the form of an Excel spreadsheet, with a template issued by the instructor, as shown in Figure 6. 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. The student (not the instructor) 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. Hands-on students were also encouraged to set up columns for example sentences from their personal corpora, as well as for common collocations (although many of the salient "words" from the vocabulary lists were collocations to begin with—for example *nominal value* in Figure 6.) [this para with Fig 6 pls]

	Word	Finance definition	Sentence from dictionary site	Translation
http://www.investopedia.com/dictionary/	capital	Money invested in a business to generate income.	The manpower and capital required to successfully market a product can be impossible for a tiny startup.	资本/Tư bản
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 sold	We expect the business to raise at least $100,000 in goodwill.	善意，友善，友好，親善…
http://dictionary.cambridge.org/dictionary/business-english/	receivables	amounts owed by customers to a company at a particular time and not yet paid	The company improved the management of its receivables by getting customers to pay faster.	应收账款
http://www.ldoceonline.com/	nominal value	Nominal value in economics also 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%, real GDP growth would be 3%.	票面价值
http://www.macmillandictionary.com/	intangible	used about something that has value for a business, although it does not exist in a physical way	Examples of intangible property include bonds, shares, copyrights, and patents.	難以捉摸的，無法形容的，難以確定的…

Figure 6 Student's vocabulary portfolio excerpt

4. RESULTS & DISCUSSION

Of the 94 students enrolled for the modules, only 55 were present for both the pre- and post-tests: 33 in the two hands-on classes, and 22 in the hands-off groups. This rather low response was because of attendance issues at the end of the semester.

Group	Pre-test mean correct score			Post-test mean correct score			Improvement (*p = 000)		
	ACC	FIN	ALL	ACC	FIN	ALL	ACC	FIN	ALL
HOn1	4.04	1.70	5.74	6.97	4.00	10.97	2.93*	2.30	5.23*
HOn2	3.80	1.54	5.34	5.73	3.29	9.02	1.93	1.75	3.68
HOff1	5.18	2.12	7.29	7.03	4.12	11.15	1.85	2.00	3.86*
HOff2	4.59	2.50	7.09	6.53	3.97	10.50	1.94*	1.47	3.41*
Hands-on groups (HOn1 & 2)	3.91	1.61	5.52	6.29	3.61	9.91	2.38*	2.00*	4.39*
Hands-off groups (HOff1 & 2)	4.89	2.30	7.19	6.79	4.05	10.83	1.9*	1.75*	3.64*

Table 2 Results from pre- and post-tests

Table 2 summarizes the results, giving mean scores in the pre- and post-tests, and the performance improvement (the difference between the two scores). Mean scores are given for each class group, as well as the combined scores for hands-on and hands-off groups. 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. The reader will notice that the scores for FIN questions are lower than those for ACC questions across the board; 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 registered post-test scores which were a significant improvement on pre-test scores, as shown in Table 2. Thus, H1 (The performances of all groups in the post-test will surpass those of the pre-test) was supported.

H2, the prediction that Hands-on DDL groups will show a greater improvement in the post-test than Hands-off groups, is also supported. Alongside that greater improvement was the unexpected finding that hands-on groups performed less well than hands-off groups on both pre- and post-tests. This appears to indicate that the hands-on groups consisted of slightly weaker students, and since the groups were assigned arbitrarily there is no immediate explanation.

HOn1 improved more on ACC questions than HOn2, while HOff1 showed less improvement than HOff2 on the same questions, so H3 was partially supported. H4 was not supported by the findings, but in the FIN case the improvement figures were not found to be statistically significant, so this finding is inconclusive. This implies that corpus work that focuses on one of two subdomains does not improve vocabulary knowledge of the one domain over that of the other. This seems counterintuitive, until one considers that the domains are closely related, and it is natural that specialist terms will appear in both. This study did not, however, consider the degree of overlap.

4.1 Results from perceptions questionnaire

In all, 60 participants out of 94 completed the post-intervention perceptions questionnaire—34 from the hands-on groups and 26 from the hands-off groups. Most of these had completed both pre- and post-tests, but as the questionnaire was administered anonymously, exact numbers are not known. Because the two groups had been subjected to different teaching approaches, a direct comparison of the success of the methods was not always possible. Participants were asked whether they found the components of the interventions "very useful", "quite useful", "not very useful" or "not useful at all". Both sets of respondents appeared to be satisfied with the different aspects of their interventions, as Table 3 illustrates.

Intervention	**Very useful/quite useful responses**			
	HOn groups		HOff groups	
	Number	%	*Number*	%
Making a corpus using files from Moodle	28	82.4	n/a	n/a

Extending the corpus to include vocabulary from the WWW	25	73.6	n/a	n/a
Studying word lists	29	85.3	22	84.6
Making a vocabulary portfolio in Excel	26	76.5	17	65.3
Doing a vocabulary quiz	27	79.4	22	84.6

Table 3 Level of participant satisfaction with interventions

A high proportion of HOn participants found the corpus creation tasks useful. The utility of directly studying a word list was around the same, according to both groups, as was that of vocabulary quizzes (this referred to the pre- and post-tests). The HOn students, however, found constructing/using the vocabulary portfolio more useful than the HOff students did. There seem to be two plausible reasons for this. The first is that they carried out the task in class, with the guidance of the instructor, while for the HOff students, working at home, the same degree of support was not available. The other possibility is that the HOn students perceived the vocabulary portfolio task as part of a meaningful workflow of which they had control from start to finish—choosing the texts for the corpus, creating and expanding the corpus, selecting vocabulary items, and finally making the portfolio. The HOff students might not have felt the same sense of task or resource "ownership" (Tyne 2009). They were told where the vocabulary lists had come from, but were not involved in their creation.

When asked whether they would continue to use the corpus and/or vocabulary portfolio resources after the end of the course, 58.8% of the HOn groups, and 57.7% of the HOff groups, claimed that they would.

5. LIMITATIONS

There were some logistical difficulties with the approach. Scheduling corpus activities into limited class time (only two hours per week) was a challenge for the researcher, especially since the overarching syllabus is determined centrally, not by individual teachers. 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.

The students were assigned to one of four weekly class events by the university timetabling team, and all the members of each group therefore constituted each of the four groups of the study. This was a matter of practicality, since it was necessary for the members of the same group to be taught together. It was not, therefore, possible to randomize the group allocations. It is not thought, however, that the timetabled groups could have any other shared characteristics, so the validity of the findings is probably not impacted by this.

A possible limitation of the study, for those considering applying the approach, is that it makes use of a commercially licensed tool, the Sketch Engine. Currently, the Sketch Engine is available at no charge to universities and other Higher Education establishments within the EU, and many universities do in any event have institutional subscriptions to the Sketch Engine (especially if they do work on lexicography or computational linguistics). For those who do not enjoy institutional access, but are able to consider applying, 30-day accounts are available, without cost or commitment, which offer all the functionality of a licensed account.

Aside from Sketch Engine, there is a variety of free or low-cost tools for creating corpora from one's own files, and most have a feature for extracting keywords (if not salient MWUs). AntConc (Anthony 2018) and WordSmith Tools (Scott 2018) are two of the most popular.

A link to the free, downloadable BootCat was given in Section 3.3.1. This can be used to seed expanded corpora from corpora originally created using a wide range of tools.

6. CONCLUSION AND FUTURE PLANS

This paper has presented a DIY corpus construction and vocabulary portfolio compilation intervention for UK university students of academic English who major in Accounting & Finance. It may be concluded that improvements in technical vocabulary knowledge were made as a result of the intervention; students reported that the experience was beneficial.

Perhaps the most significant finding of this study is that performance within both hands-on and hands-off groups (between the pre- and post-tests) improved. It is clear that the use of the DDL approach helped both groups of learners. In terms of the between-groups performance, the hands-on

groups appeared to have benefited to some extent from the corpus construction task, since their performance improvement was slightly better than the hands-off group, although the evidence for this is less compelling. It is known from attendance figures (including the numbers of participants in the pre- and post-tests, and questionnaire), and from tutors' own (anecdotal) observations that the HOn students were more engaged, and we find this encouraging (particularly in view of the fact that HOn students were apparently, as reported in the Results and Discussion, of slightly lower proficiency than HOff).

It will be interesting to see if these results generalize to larger student numbers; our inter-departmental collaborations mean that in future cohorts, it will be possible to repeat the experiment on a larger scale. Future cohorts majoring in International Business and Engineering, for example, will be asked to create corpora and vocabulary portfolios as part of their Academic English modules. Cohorts other than Accounting & Finance are more mixed in terms of L1 background, allowing us to study demographic patterns arising from the hands-on/off DDL interventions. Almost all the participants in this study were from China, so the findings reported here may be a function of learning styles associated with that particular culture.

The corpus-informed lexical resource creation tasks described here provide a motivating, meaningful and workflow-driven way for students to access and learn the terminology and usages of their own specialist subjects. It is hoped that readers will be inspired to try out corpus construction and vocabulary portfolio creation with their own Academic English students.

REFERENCES

Ackermann, K. and Chen, Y. H. (2013). Developing the Academic Collocation List (ACL): A Corpus-driven and Expert-judged Approach. *Journal of English for Academic Purposes*, *12*(4), 235-247.

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.

Anthony, L. (2018). AntConc (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

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 July 1, 2019, from http://www.sslmit.unibo.it/~guy/graz.htm

Author (2011)

Author (2015a)

Author (2015b)

Baroni, M. & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the Web. In *Proceedings of 4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal, 1313–1316. Retrieved July 1, 2019, from https://pdfs.semanticscholar.org/a1ea/c69123e1acbe3f248e1ce85e94ae67b0bbe8.pdf

Baroni, M., Kilgarriff, A., Pomikálek, 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 July 1, 2019, from http://www.mt-archive.info/EAMT-2006-Baroni.pdf

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.

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

Boulton, A. (2008). 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 July 1, 2019, from http://hal.archives-ouvertes.fr/docs/00/32/69/80/PDF/XXXX_boulton_TaLC_interdisciplinary.pdf

Browne, C., Culligan, B., & Phillips, J. (2013). *A New Academic Word List*. Retrieved July 1, 2019, from http://www.newgeneralservicelist.org/nawl-new-academic-word-list

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

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 1, 2019, from http://wackybook.sslmit.unibo.it/pdfs/castagnoli.pdf.

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.

College Entrance Examination Center. (2002). 大學入學考試中心高中英文參考詞彙表 [High School English Reference Wordlist]. Retrieved July 1, 2019, from http://www.ceec.edu.tw/Research/paper_doc/ce37/2.pdf

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

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

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.

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

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

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.

Hyland, K. and Tse, P. (2007). Is there an "Academic Vocabulary"? *TESOL Quarterly, 41*(2): 235-253.

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.

Johns, T. (1986). Micro-Concord: a language learner's research tool. *System, 14*(2): 151–162.

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.

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 July 1, 2019, from http://www.kilgarriff.co.uk/Publications/2008-KilgEtAl-euralex-gdex.doc

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In *EURALEX 2004 Proceedings*, Lorient, France. Retrieved July 1, 2019, from http://kilgarriff.co.uk/Publications/2004-KilgRychlySmrzTugwell-SkEEuralex.rtf

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. Retrieved July 1, 2019, from http://webcache.googleusercontent.com/search?q=cache:http://ict4lt.org/en/en_mod2-4.htm

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.

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

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

Neufeld, S. (2012). *Data driven language learning: Using corpus linguistics in language learning*. Retrieved July 1, 2019, from http://www.scoop.it/t/data-driven-language-learning

Oxford Reference (2016) *Oxford Reference*. Retrieved July 1, 2019, from http://www.oxfordreference.com/

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

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

Rychlý, P. (2008). A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing*, RASLAN, pp. 6–9.

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

Scott, M. (2018). WordSmith Tools version 7, Stroud: Lexical Analysis Software.

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

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

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

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: Mélanges CRAPEL. Retrieved July 1, 2019, from https://hal.archives-ouvertes.fr/hal-00416544/document

University of Michigan (2011). *MICASE Kibbitzers*. Retrieved July 1, 2019, from https://web.archive.org/web/20111008033810/http://micase.elicorpora.info/micase-kibbitzers

Vincent, B. (2013). Investigating academic phraseology through combinations of very frequent words: a methodological exploration. *Journal of English for Academic Purposes 12*(1), 44-56.

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

Wang M., & Nation P. (2004) Word meaning in academic English: Homography in the academic word list. *Applied Linguistics 25*(3), 291-314

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

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

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

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