Taiwan Journal of TESOL
Vol. 12.2, 117-141, 2015

THE DEVELOPMENT OF A CORPUS-BASED TOOL FOR EXPLORING DOMAIN-SPECIFIC COLLOCATIONAL KNOWLEDGE IN ENGLISH

Ping-Yu Huang, Chien-Ming Chen, Nai-Lung Tsao, & David Wible

ABSTRACT

Since it was published, Coxhead's (2000) Academic Word List (AWL) has been frequently used in English for academic purposes (EAP) classrooms, included in numerous teaching materials, and re-examined in light of various domain-specific corpora. Although well-received, the AWL has been criticized for ignoring some important facts that words still tend to show irregular distributions and are used in different ways across disciplines (Hyland & Tse, 2007). One such difference concerns collocations. Academic words (e.g. *analyze* and *concept*) often co-occur with different words across domains and sometimes even refer to different meanings. What EAP students need, accordingly, is a "discipline-based lexical repertoire" (Hyland & Tse, p.235). Inspired by Hyland & Tse's insightful remarks, we developed an online corpus-based tool, *TechCollo*, which is meant for EAP students to explore collocational knowledge in a domain or compare collocations across disciplines. TechCollo runs on textual data stored in three specialized corpora and utilizes frequency and some information-theoretical measures (e.g. mutual information) to decide whether co-occurring word pairs constitute collocations. In this article we describe the current version of TechCollo and how to use it in EAP studies. Particularly, we report a pilot study in which we employed TechCollo to investigate whether the AWL words take different collocates in different domain-specific corpora. This pilot basically confirmed Hyland & Tse's indications and demonstrated that many AWL words show uneven distributions and collocational differences across disciplines.

Key Words: domain-specific collocations, domain-specific corpora, online learning tool, English for academic purposes

INTRODUCTION

In EFL (English as a foreign language) contexts, it has been widely acknowledged that vocabulary constitutes a crucial, if not the most challenging part of English learning. For students who learn English for academic purposes (EAP) in such contexts, the vocabulary learning task that they face seems even more complex or difficult than that of English for general purposes (EGP) students. EGP students basically have to memorize thousands of words which are frequent in common use. EAP students, however, need to learn vastly more words including both technical terms as well as so-called academic words which appear more often in academic discourse than in other types of texts.[1] To help EAP students become familiar with academic vocabulary, several researchers thus have attempted to collect words worth focusing on in EAP courses. Among those attempts, Coxhead's (2000) Academic Word List (AWL) has been generally considered the most complete and successful work. To create the AWL, specifically, Coxhead identified 570 word families which appeared to be *specialized* in academic discourse and *generalized* across four professional domains: law, arts, commerce, and science. Coxhead removed the most frequent words of English from analysis and used simple statistics to decide on words which were particularly frequent in academic texts. Collectively, the 570 word families made up around 10% of Coxhead's academic corpus, and only 1.4% of a corpus of English fiction. One of the main reasons for the wide acceptance of the AWL is that it enables instructors to set an achievable goal for EAP students. The AWL, furthermore, was divided into ten sublists based on word frequency, giving EAP teachers a useful guide to lesson planning during students' study progress.

Although well-received, the AWL is not without criticism. Chen and Ge (2007), for instance, pointed out that the AWL words were too general to be addressed to medical students. Among its 570 word families, only 51.2% of them were found to be frequent in medical journal articles. Similar problems have been reported and discussed by Vongpumivitch, Huang, and Chang (2009) and Martínez, Beck, and

[1] Academic words have also been termed *sub-technical vocabulary* (Yang, 1986), *semi-technical vocabulary* (Farrell, 1990), and *specialized non-technical lexis* (Cohen, Glasman, Rosenbaum-Cohen, Ferrara, & Fine, 1979) in the literature. In this article, these terms are used interchangeably to refer to lexical items which appear more often in academic texts than in other types of texts.

Panza (2009) using applied linguistic or agricultural texts. Arguably the strongest criticism of the AWL came from Hyland and Tse (2007), who questioned the design and analysis of Coxhead's (2000) study and even the real existence of an academic literacy. That is, Hyland and Tse doubted whether there was a single set of words which "represented the vocabulary of academic discourse" and were "valuable to all students irrespective of their field of study" (p. 238). The AWL, though covering certain proportions of texts across disciplines, still tends to show better coverage of texts in some areas (e.g. social sciences) than in others (e.g. anatomy) (Chung & Nation, 2004). Such uneven coverage suggests that, for students in areas such as anatomy, studying the AWL will "involve considerable learning effort with little return" (Hyland & Tse, p. 236). Furthermore, as Hyland & Tse argued, traditional word lists created for academic purposes tend to ignore the important fact that words appearing in different domains are embedded in different phraseological patterns. Those phraseological differences sometimes even lead to semantic differences. Vocabulary lists which fail to highlight such discipline-specific meanings and usages may give students a misleading impression that words are used in similar ways in different disciplines. Hyland and Tse's criticisms and discussions remind us that what EAP students need is "a more restricted, discipline-based lexical repertoire" (p.235), rather than a core vocabulary list for students no matter which domain they are in.

Inspired by Hyland and Tse's (2007) insightful remarks, we developed an online corpus-based tool which enables EAP students to explore and study domain-specific lexical knowledge that they need for academic purposes. The knowledge that we concentrate on concerns collocational patterns. Knowing a word, according to the Firthian view of vocabulary, involves knowing other words commonly appearing around it. Frequent and common word combinations, rather than single lexical items, represent the discipline-specific knowledge which EAP students must acquire in order to correctly figure out technical meanings of words in their own domain(s). Our online tool, which is called TechCollo (i.e. technical collocations), is meant to be used by EAP students to search for and explore discipline-specific collocations. It runs on textual data stored in medium-sized domain-specific corpora, each containing millions of running tokens. With TechCollo, EAP students can easily check whether a two-word combination is common in their domain, differentiate lexical usages in two different domains, and, by comparing collocational patterns in a specialized and a general-purpose corpora, investigate whether some usages are restricted to certain domain(s). A good example of the domain-specific collocations discussed here is the word *proceeding* used in legal studies. Whereas *proceeding* often refers to a collection of research papers in contemporary academic societies, it means a legal action and frequently goes with verbs such as *file* and *conduct* only in legal texts. TechCollo, with its ability to run and process lexico-grammatical knowledge in domain-specific corpora, will provide such specialized lexical and collocational knowledge for EAP students.

The remainder of this article is organized as follows. First, we review and discuss traditional core academic word lists such as the AWL. We intend to more thoroughly review their contributions and criticisms. Next, a detailed description of our tool is offered, including the technical corpora underlying it, extractions of collocations, user interfaces, and some main functions. To specify how to use TechCollo to explore cross-disciplinary differences, we then present the findings of a pilot study within which we examined whether the AWL words tend to show collocational differences in different domains. We conclude our paper by discussing some future plans for improving TechCollo.

STUDIES ON ACADEMIC VOCABULARY

Academic words, defined as “formal, context-independent words with a high frequency and/or wide range of occurrence across scientific disciplines, not usually found in basic general English courses” (Farrell, 1990, p. 11), are regarded as non-salient expressions due to their supportive rather than central role in written texts (Coxhead, 2000). Non-salient as they are, academic words are claimed to cover considerable proportions of words in texts across fields of study and to be important knowledge that EAP students must master in order to succeed in their undergraduate or postgraduate studies. In the research literature, some early studies were conducted to identify/analyze academic vocabulary which either assumed that students had already learned general service words, and they investigated, in addition to those words, what words appeared across different disciplines (e.g. Campion & Elley, 1971) or collected the words that foreign students of English found difficult and wrote annotations or translations in their university textbooks (e.g. Ghadessy, 1979). The two different approaches, however, were found to generate word lists with substantial overlaps (Nation, 2001). Xue and Nation (1984) combined some of those early lists into one complete collection, the University Word List (UWL). The UWL in total contained 836 word families, showing a rather impressive coverage of 8-9% of academic texts. In non-academic texts such as newspapers, the UWL covered less than 5%, which suggested the academic nature of the list.

Although the UWL was widely used for a number of years, it was problematic in that it was not established based on consistent selection criteria, and the corpora or studies which it considered were not well-balanced. To develop a list which drew upon selection principles from corpus linguistics and better represented texts in a variety of professional domains, Coxhead (2000) collected texts coming from four important fields of study: arts, commerce, law, and science, and compiled the Academic Word List. Specifically, Coxhead's corpus contained 3.5 million running tokens, which were evenly distributed in the four disciplines. The corpus was composed of 414 texts, including articles taken from professional journals, university textbooks, lab manuals, as well as academic sections of some large-scale corpora such as the Lancaster-Oslo/Bergen Corpus (Johansson, 1985). Like the UWL, Coxhead used word families as basic entries for the AWL. The focus on word families, as Coxhead explained, solved the problem that words might be morphologically distinct but semantically closely related (e.g. the words *dimension* and *multidimensional* listed on Sublist 4 of the AWL). Additionally, to count or define word families as main entries was supported by research showing that members of a word family were connected together in the mental lexicon (Nagy, Anderson, Schommer, Scott, & Stallman, 1989). Concerning word selection, four principles were applied to decide what to include into the AWL:

1. Exclusion of general service words: The most widely used 2,000 words, as described by West's (1953) General Service List (GSL), were removed;
2. Frequency: Members of a word family had to appear at least 100 times in the academic corpus;
3. Uniformity: A member of a word family had to occur over 10 times in each of the four disciplines;
4. Range: A member of a word family had to appear in at least 15 of the total 28 subject areas.

Based on the four principles, Coxhead identified 570 word families which accounted for around 10% of her academic corpus and covered 9.1-12% of the four sub-corpora. Together with the GSL, the two lists showed coverage of 86% of the entire database. Evaluated with tokens in a corpus of fiction, the AWL accounted for only 1.4% of its total words, a percentage confirming that the AWL words were specific to academic discourse. In addition to these impressive results, Coxhead compared the AWL with the UWL, demonstrating that the AWL, with far fewer entries, showed slightly higher coverage (10% vs. 9.8%). To be easily applied to classroom teaching, furthermore, the AWL words were categorized into ten sublists, each containing 60 word families and the last including 30. Since it was published, the AWL has been frequently used by teachers in EAP courses, covered by a great number of teaching materials, and re-examined by applied linguists using various domain-specific corpora. The AWL, as Coxhead (2011) herself claims, indeed has exerted a much greater influence than the author ever imagined.

As stated earlier, despite the wide acceptance and popularity, the AWL has still been criticized by some researchers concerning its design, assumptions, coverage, and usefulness. Here we focus on Hyland and Tse (2007) which arguably has been the most thorough examination of the AWL so far and directly inspires the present study. According to Hyland and Tse, the AWL, as well as other similar cross-disciplinary word lists, inevitably suffers from two important weaknesses. The first concerns their assumptions and the second is about the phraseological behaviors of lexical items. First, concerning assumptions, as Hyland and Tse pointed out, for the compilation of academic word lists to be reasonable or valid, there should be a core lexical repertoire which is shared by all professional areas and valuable to all EAP students. However, previous academic word lists, including the AWL, appear to ignore the fact that words very often behave differently in terms of frequency and range. Hyland and Tse criticized that the design of Coxhead's corpus was rather "opportunistic" (p. 239) and did not contain equal numbers of texts across disciplines. To reveal a clearer picture of the distribution of the AWL in different domains, Hyland and Tse created an academic corpus which was better controlled and included equal numbers of texts representing three selected disciplines: sciences, engineering, and social sciences. This new academic corpus, as Hyland and Tse demonstrated, generated results markedly different from those of Coxhead. In the three selected domains, the AWL together with the GSL accounted for only 78.3% of the tokens in sciences, 10% less than that in social sciences. The 10% difference suggested that sciences students would find the AWL (plus the GSL) less valuable or useful than social sciences majors. Hyland & Tse further applied certain principles to identify frequent words in their database. They focused on the items whose occurrences were higher than the average for all AWL words. This resulted in the inclusion of only 192 word families, which could be considered frequent in Hyland and Tse's corpus. A very interesting finding observed by the authors was that their 60 most frequent families and those found by Coxhead were substantially different. This finding once again confirmed the difficulty of establishing a complete word list for EAP purposes.

The second weakness discussed by Hyland and Tse (2007) was about the semantic and collocational properties of words. As their corpus data revealed, words tended to show meaning and phraseological differences across fields. The word *process*, according to Hyland and Tse, was more likely to appear as a noun in science than in social sciences. Words such as *value*, though frequent in many different areas, took on specialized meanings as it co-occurred with *stream* in computer science. While describing a particular point during an activity, different words were favored by researchers in different domains, including: *phase* in biology, *stage* in mechanical engineering, and *period* in applied linguistics. The consideration of only single lexical items, as the authors argued, not only ignored important discipline-specific usages, but caused a misunderstanding that words were used in similar ways across domains. Although the AWL has its pedagogical merits, it failed to recognize current conceptions of EAP that different academic areas "constitute a variety of subject-specific literacies" (p. 247).[2]

[2] Although Hyland and Tse's (2007) research clearly presents problems and severe limitations of academic vocabulary lists, some applied linguists seem to ignore those problems and continue to create core word lists for EAP purposes. One such list appears in a recent publication by Gardner and Davies (2013), who compiled the Academic Vocabulary List (AVL) based on a 120-million-word academic part of the Corpus of Contemporary American English (COCA). Using more sophisticated frequency and dispersion statistics, Gardner & Davies collected a total of 3,000 lemmas frequent in academic discourse. This new list, compared with the AWL, is found to show even higher coverage of academic texts (13.7% vs. 6.9% of the academic sub-corpus of British National Corpus). Although we acknowledge that it is likely to keep improving the coverage with better statistical techniques as what Gardner and Davies do, we believe that general academic vocabulary collections inevitably suffer from the problems that

The study reported in this article aims to devise and develop a learning tool which is able to generate the lexico-grammatical knowledge that EAP students actually need. Rather than extracting common and shared words from academic corpora, we create some domain-specific corpora and intend to collect the knowledge which is particularly useful and important for students in the selected domains. What our learning tool offers is the domain-specific lexical/collocational knowledge that EAP students can study to figure out the phraseological and semantic behaviors of words. Such knowledge, based on Hyland and Tse's (2007) analysis and comments, will enable students to familiarize themselves with the real usages of words in their own domain(s).

TECHCOLLO: DEVELOPMENT AND MAIN FUNCTIONS

To develop a tool which is able to offer technical collocations, first we consulted research in the fields of electronic lexicography and automatic term recognition (ATR). Our intentions were to understand whether and how previous researchers in those areas retrieved multiword terminology from texts. Basically, it was found that certain frequency-based and statistics-based measures (e.g. t test and log-likelihood test) have been utilized to determine the termhood of word combinations. Some ATR studies, however, took more sophisticated extraction approaches. For instance, Wermter and Hahn (2005) distinguished domain-specific from non-domain-specific multiword items on the basis of word strings' *paradigmatic modifiability* degrees which were assumed to be lower for domain-specific strings.

A corpus study worth noting here was Barrière (2009), in which the author also aimed to help students learn domain-specific collocations online. Specifically, Barrière established an online platform, TerminoWeb, for foreign language learners to upload technical articles. Learners were guided to select unknown terms in those articles and the platform also automatically identified certain terms. Next, a set of

Hyland & Tse pointed out. Take the word *claim* (as a noun) listed among the top 500 lemmas in the AVL as an example. In the medical and legal corpora that we constructed for TechCollo, *claim* shows remarkably uneven distributions in the two datasets (40 vs. 428 per million words). Such cross-disciplinary variations are even more apparent and interesting when we consider its collocates in the two disciplines (e.g. *claim* tends to refer to "an official request" when it collocates with *adjudicate* in law.)

queries were performed on the Web to collect texts relevant to the source text(s) (i.e. passages including the selected and identified terms). The collected texts or webpages were then a domain-specific corpus. Within it, users could conduct concordance searches to understand the meanings of a term or make collocation searches for the term. The calculation of collocations performed by Barrière (2009) was based on Smadja's (1993) algorithm, which, as Smadja claimed, reached a precision rate of 80% for collocation extraction.

Although both TerminoWeb and our TechCollo were meant to be used by students to learn domain-specific collocations online, there are some key differences between the two platforms. First, unlike the technical corpora which were compiled via the TerminoWeb with texts from the whole Web and likely to include lots of messy data, the corpora underlying TechCollo were composed of texts edited in advance which are cleaner and more reliable. TechCollo, furthermore, offers an interface which allows users to compare collocations in two different specialized domains or in a specialized and a general-purpose corpus. These convenient search functions will more effectively enable EAP learners to discover and explore specialized collocational knowledge online.

To illustrate the main functions of TechCollo, below we respectively describe: (1) the compilation of domain-specific corpora underlying it, (2) the determination or identification of a word pair as a candidate for a true collocation, and (3) the interface designed for EAP students.

Corpora Underlying TechCollo

The current (and the first) version of TechCollo extracts collocations from three domain-specific corpora. All of them are medium-sized databases, containing 4.8-7.9 million running tokens. Each of them is composed of texts coming from the largest online encyclopedia, Wikipedia, as well as from high-quality journal articles. The Wikipedia texts that we used and processed were provided by the WaCky team of linguists and information technology specialists (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009),[3] who compiled huge Wikipedia corpora

[3] The Wikipedia corpus that we downloaded from the Wacky website (http://wacky.sslmit.unibo.it/) was WaCkypedia_EN, which was POS-tagged, lemmatized, and syntactically parsed with TreeTagger and MaltParser. We thank Baroni et al. (2009) for offering the WaCkypedia_EN corpus.

for various European languages such as English, Italian, and French. Based on an English Wikipedia corpus created by the WaCky team, we built up corpora for three domains: medicine, engineering, and law. The second source was writings from journal articles. That is, for the same medical, engineering, and legal domains, we consulted more than sixty top academic journals and respectively downloaded 280, 408, and 106 papers from certain academic journal websites online. We utilized the tools offered by Stanford CoreNLP (Klein & Manning, 2003) to POS-tag and parse those journal texts and then added them to the medical, engineering, and legal corpora. The sizes of our three specialized corpora are shown in Table 1. In addition to the domain-specific databases, TechCollo also provides collocation searches in a general-purpose corpus: British National Corpus (BNC). We offer collocation exploration for this large-scale non-domain-specific database in order for users to compare and study collocations in subject areas and general use. Table 1 also provides the number of tokens included in the BNC.

Table 1

Corpora Underlying TechCollo

Corpus	Token Count	From Wikipedia	From Journals
Medicine	4,858,579	2,812,082	2,046,497
Engineering	5,840,931	3,706,525	2,134,406
Law	7,947,895	5,556,661	2,391,234

Collocation Extraction

Various measures have been employed in computational linguistics to automatically identify collocations in texts. Those measures can be roughly divided into three categories (Wermter & Hahn, 2004): (1) frequency-based measures, (2) information-theoretical measures (e.g. mutual information), and (3) statistical measures (e.g. t test and log-likelihood test). To evaluate whether a measure is effective or to compare the effectiveness of several measures, one often needs to collect a set of true collocations and non-collocations and examine how a measure ranks those word combinations (see, for example, Pecina, 2008).

An important lesson learned from the examinations of those measures is that there is no single measure which is perfect in all situations. To identify target collocations, it is suggested that one has to exploit several association measures with a correct understanding of their notions and behaviors.

TechCollo utilizes three main measures to decide whether a two-word combination constitutes a good candidate collocation in a five-word window in our textual databases: frequency, traditional mutual information (tradMI) (Church & Hanks, 1990), and normalized mutual information (normMI) (Wible, Kuo & Tsao, 2004). A learner using TechCollo can set or change values of these measures to show candidate collocations in our three technical corpora. First, frequency refers to raw co-occurrence count of a word pair. However, to filter out pairs which are extremely frequent as a result of one or both of their component words but are not true collocations,[4] TechCollo offers the common association measure, tradMI, which is formulated as follows.

$$tradMI(x,y) = \log_2 \frac{P(x,y)^2}{P(x)\,P(y)}$$

This information-theoretical measure works by comparing the joint probability of two expressions *x* and *y* (i.e. the probability of two expressions appearing together) with the independent probabilities of *x* and *y*. In other words, tradMI expresses to what extent the observed frequency of a combination differs from expected. Although tradMI effectively removes word pairs containing high-frequency words, it inevitably suffers from a problem that it also filters out certain pairs which contain high-frequency words but are interesting and actual collocations. In English, for example, word combinations such as *take medicine, make (a) decision, and run (a) risk* are real collocations which include very frequent component words. To solve the problem with the tradMI, Wible, et al. (2004) introduced the alternative association measure: normMI, which attempts to minimize the effects caused by sheer high frequency words. To achieve this, Wible et al. normalized the

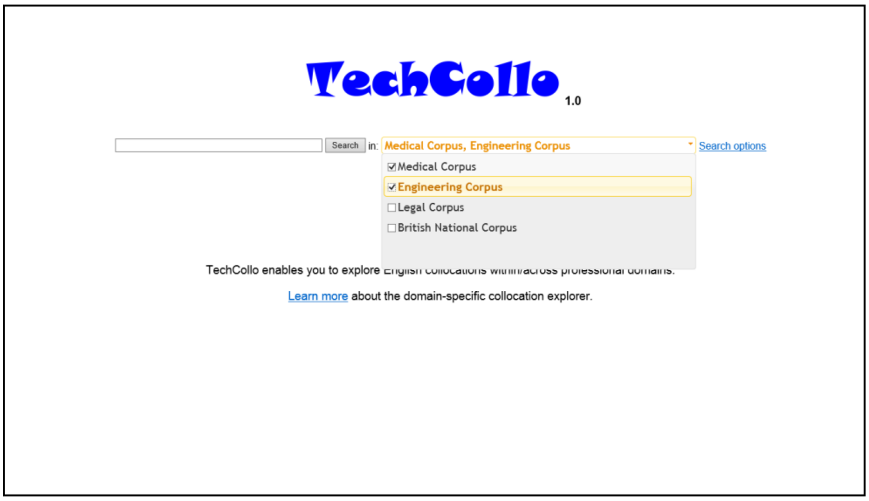
[4] A typical example of frequent non-collocational pairs is the string *of the*, which appears more than 2.7 million times in COCA (Davies, 2008).

tradMI by dividing the lexeme frequency by its number of senses (based on WordNet). The formula for the normMI is shown below. Basically, the notion of normMI is based on the *one sense per collocation* assumption proposed by Yarowsky (1995). A highly frequent word (e.g. *take, make,* and *run*) is generally polysemous. However, as Yarowsky indicated, as the word appears in a collocation, it is very common that only one of its senses is used (e.g. the word *run* in the collocation *run a risk*). Wible et al. compared tradMI with normMI using several pairs containing high-frequency words (e.g. *make effort* and *make decision*) and found that these combinations are ranked higher among the identified candidate collocations by normMI. It is important to note, although the normMI produces higher recall than the tradMI, precision does not decrease accordingly. On our TechCollo interface, we provide the normMI and expect that EAP learners can use it to find and learn certain word combinations which include highly frequent words but are still true and specialized collocations in their domain(s).

$$normMI(x,y) = \log_2 \frac{P(x,y)^2}{\left({}^{P(x)}\!/_{sn(x)} \right) * \left({}^{P(y)}\!/_{sn(y)} \right)}$$

User Interface

The main page of TechCollo is shown in Figure 1. Basically, this online collocation exploration tool allows users to choose from three medium-sized domain-specific corpora: medical, engineering and legal corpora, and a general-purpose corpus: BNC. A user accessing the website can key in a keyword that he/she intends to study and the system will automatically search for words which tend to co-occur with the keyword in the selected databases. The current released version of TechCollo provides searches of verb-noun collocations. The values of frequency and tradMI, as specified earlier, can be changed and decided by users so that the system will respond with either a shorter list of word pairs with higher frequency counts and tradMI or a longer list containing more candidate collocations.



Medical Corpus, Engineering Corpus

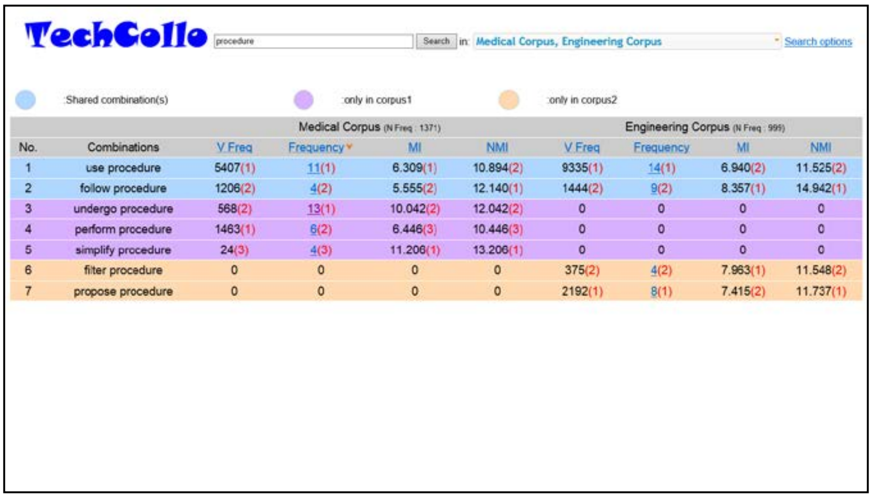
☑ Medical Corpus
☑ Engineering Corpus
□ Legal Corpus
□ British National Corpus

TechCollo enables you to explore English collocations within/across professional domains.

Learn more about the domain-specific collocation explorer.

Figure 1. Main page of TechCollo

Here we take the noun *procedure* and its verb collocates in medical and engineering corpora as examples. We fed this word into the TechCollo system with the frequency and tradMI set at 3 and 4, respectively. That is, only the verbs which appear together with *procedure* at least four times and having mutual information larger than 4 were identified as candidate collocates. The search results are shown in Figure 2.



● :Shared combination(s) ● :only in corpus1 ● :only in corpus2

		Medical Corpus (N Freq : 1371)				Engineering Corpus (N Freq : 999)			
No.	Combinations	V Freq	Frequency	MI	NMI	V Freq	Frequency	MI	NMI
1	use procedure	5407(1)	11(1)	6.309(1)	10.894(2)	9335(1)	14(1)	6.940(2)	11.525(2)
2	follow procedure	1206(2)	4(2)	5.555(2)	12.140(1)	1444(2)	9(2)	8.357(1)	14.942(1)
3	undergo procedure	568(2)	13(1)	10.042(2)	12.042(2)	0	0	0	0
4	perform procedure	1463(1)	6(2)	6.446(3)	10.446(3)	0	0	0	0
5	simplify procedure	24(3)	4(3)	11.206(1)	13.206(1)	0	0	0	0
6	filter procedure	0	0	0	0	375(2)	4(2)	7.963(1)	11.548(2)
7	propose procedure	0	0	0	0	2192(1)	8(1)	7.415(2)	11.737(1)

Figure 2. Search results for the word *procedure*

According to the results offered by TechCollo, there were, respectively, 1371 and 999 tokens of *procedure* in the two technical corpora. The two corpora (or the two fields of profession) shared two common collocations: *use procedure* and *follow procedure*. Taking a closer look at the *unshared* verb collocates which appeared in medicine but not in engineering, we found that *procedure* tends to co-occur with *undergo* only in medicine. The example sentences (which can be accessed by clicking on the frequency numbers on the interface) for the "*undergo procedure*" pair informed us that *procedure* is a technical term in medicine which refers to a medical operation. With the rich corpus resources available on TechCollo, we expect and encourage EAP students to discover such specialized collocations by: (1) searching collocations in a specific domain, (2) comparing collocations in two domain-specific corpora (e.g. medical vs. engineering corpora), and (3) comparing collocations in a specialized and a general-purpose corpora (e.g. medical corpus vs. BNC).

On TechCollo, for the extracted candidate collocations, a user can change their ordering(s) by clicking on the icons "frequency" or "MI" (which refers to tradMI). The other measure shown on the TechCollo interface is NMI, which is the normMI that we described earlier and provide on our website in the hope that it allows EAP learners to find certain true collocations containing high-frequency component words. To examine the effectiveness of the normMI, we tested it with the legal collocation "*break law*." In our legal corpus, if we used tradMI to search for verb collocates for the noun *law*, there was a great possibility that *break* would not be noticed because it was highly frequent and the pair *break law* was thus ranked 48th by tradMI among the extracted verbs. NormMI, however, successfully changed the ranking of the pair which was ranked in a much higher position (i.e. the 8th). This example gave us good evidence that the normMI is useful for raising collocations containing high-frequency words in more advantaged positions for learners to pay attention to them. A more thorough examination, nevertheless, is required to investigate whether the normMI is indeed an effective measure of identifying collocations in domain-specific texts.

COMPARISON OF COLLOCATIONS ACROSS DOMAINS: A PILOT STUDY

To specify and illustrate how to use TechCollo in EAP study, we ran a pilot study within which we examined the verb-noun collocations in two different domains: medicine and engineering. More specifically, we focused on the nouns included in the Sublist 1 of the AWL (Coxhead, 2000) and explored and analyzed their verb collocates in our medical and engineering corpora. Our purpose, then, was to verify whether it is true that words tend to show differences in phraseological behaviors in different professional areas, as Hyland and Tse (2007) pointed out.

First, from the sixty word families contained in the Sublist 1, we identified 109 nouns. Those nouns were fed into TechCollo in order to extract their frequent co-occurring verbs in the medical and engineering corpora. In this pilot, the frequency and tradMI were set at 1 and 3, respectively; that is, only the verb-noun combinations which appeared at least two times and took mutual information scores larger than 3 were identified as collocation candidates. In the data generated by TechCollo, the very first observation that we made was that at least 20% of the 109 nouns showed rather uneven distributions in the two selected databases. Some examples of those nouns are demonstrated in Table 2. We can find in this table that several AWL nouns show up rather frequently in one domain but not in the other (e.g. *policy, principle,* and *sector*). Some other nouns, though being quite common and appearing more than one million times in both disciplines, were around three or four times more frequent in a corpus than the other (e.g. *individual, method,* and *role*).

These distributional variations suggest that an academic word which is highly frequent and important in one discipline may be less common and important for students in another (e.g. the words *sector* and *specification* for medical school students). EAP students who are required to study the AWL for their academic studies are very likely to run the risk of being exposed to more lexical items than they actually need.

Table 2

Nouns with Highly Irregular Distributions in Medical and Engineering Corpora

Word	Frequency (per million tokens) in Medicine	Frequency (per million tokens) in Engineering
assessment	160	65
concept	105	293
distribution	140	415
evidence	472	89
individual	389	111
issue	119	262
majority	122	58
method	572	1107
policy	37	125
principle	77	190
processing	98	350
requirement	77	318
response	843	272
role	696	184
section	152	517
sector	7	99
specification	14	136
specificity	77	5
theory	172	383
variant	119	48

In addition to the comparisons of numbers of occurrence of the nouns, what interests us more concerns their relations with verbs in medicine and engineering. We present part of the verb-noun collocation data in Table 3.

Table 3

Verb Collocates in Medical and Engineering Corpora

Noun	Shared Collocates	Verbs in Medicine Only	Verbs in Engineering Only
analysis	*perform, conduct*	*undertake, run*	*carry out*
approach	*adopt, develop*	*provide*	*propose, suggest*
area	*select, affect*	*identify, depict, delineate*	*consider*
assumption	*violate*	*assess, evaluate*	*estimate, contradict*
availability	*increase*		
benefit	*provide, offer*	*confer*	*achieve, produce*
concept	*suggest*		*propose*
data	*analyze, collect, obtain, evaluate*	*interpret, conflict, contradict*	*deal, handle*
definition	*use*	*propose*	*introduce*
distribution	*show, estimate, illustrate*	*depict, examine, display*	*measure, evaluate, exhibit*
estimate	*provide, yield, result in*	*generate, construct*	*form, lead to*
estimation		*reach*	*achieve*
evidence	*present, show, detect*	*confirm, accumulate, examine*	*generate*
formula		*fortify*	*derive*
function	*modify*	*impair, inhibit, disrupt*	*describe, manipulate*
issue	*address, raise, clarify*	*explore*	*deal with*

Table 3

Verb Collocates in Medical and Engineering Corpora (continued)

Noun	Shared Collocates	Verbs in Medicine Only	Verbs in Engineering Only
method	*develop, propose, apply*	*perform, utilize*	*employ*
occurrence		*decrease*	*avoid*
percentage	*show, increase*		
period	*consider*		
Noun	*Shared Collocates*	*Verbs in Medicine Only*	*Verbs in Engineering Only*
principle	*underlie*	*adhere*	*follow*
procedure	*follow*	*undergo, receive*	*employ*
process	*affect, influence*		
processing		*undergo*	*implement, conduct*
requirement	*meet, fulfill*		*satisfy, achieve*
research	*perform, conduct, undertake*		*carry out*
response	*elicit, investigate*	*induce, activate, trigger, boost*	*amplify*
role	*play, explain, examine*	*assess, evaluate*	
section	*discuss*	*stain*	*explain, mention, provide*
significance		*achieve, reach, assess*	*investigate*
structure		*detect*	*analyze, estimate*
theory		*propose, confirm*	*develop, formulate*
variable	*determine*		
variation	*exhibit, observe*	*reflect*	

As Table 3 displays, there are several nouns which share verb collocates in the medical and engineering corpora, including: *availability, percentage, period,* and *variable*. In other words, these verb-noun combinations are of almost equal importance for EAP students, at least for medicine and engineering majors. This table, however, reveals that there are many more so-called generalized academic words which tend to take different collocates and some of them even refer to different meanings across disciplines. The word *area*, for example, co-occurs with *identify, depict,* and *delineate* only in medicine and refers to the specialized meaning of a part on the surface of human body. Some other nouns, such as *function, procedure,* and *response* also contain such medicine-specific senses as they co-occur with *impair, undergo,* and *activate*, respectively. The word *formula*, more interestingly, has two different meanings: (1) amounts of ingredients for making something and (2) numbers, letters, and symbols that represent a mathematical rule as it takes *fortify* or *derive* in medicine or in engineering.

Perhaps the most notable cross-disciplinary difference shown by the collected collocations is, while expressing a similar idea, people in medicine and engineering seem to prefer different verbs. According to the data in Table 3, at least twenty nouns demonstrate such variations. Examples for this phenomenon are: *undertake/carry out analysis, assess/estimate assumption, propose/introduce definition, display/exhibit distribution, construct/form estimate, reach/achieve estimation, utilize/employ method, adhere/follow principle, assess/investigate significance*, etc. These field-specific idiomatic and habitual usages do not suggest that they are used only in one discipline and not in another. Rather, they provide evidence showing that people in different areas indeed tend to select different word combinations which form their particular and domain-specific literacies (Hyland & Tse, 2007). What EAP students need to study, accordingly, should be these common specialized collocations and usages which make their writings and speech professional in their own domain(s).

CONCLUSION

The pilot study reported in this article clearly suggests that academic words, though being collected for EAP students irrespective of their subject areas, tend to have different numbers of occurrence and co-occur with different words in different domains. If students depend on word lists such as Coxhead's (2000) AWL or Gardner and Davies's (2013) Academic Vocabulary List to learn academic words, they are very likely to memorize more lexical items than they actually need for studies in their own domain(s). Plus they will not be familiar with the common and important collocations that their colleagues frequently use in speech or writing. What EAP students need, or more specifically, what EAP researchers are suggested to collect, should be discipline-based and discipline-specific vocabulary and collocation knowledge. To generate and offer such resources, we develop the online corpus-based collocation exploration tool, TechCollo, with the aim of providing the specialized lexico-grammatical knowledge that EAP students need to master at college. Our tool, with its ability to allow students to check specialized collocations in a discipline, differentiate collocations across disciplines, and compare collocations in domain-specific and general-purpose corpora, is of great help for EAP students who would like to know word usages when they learn to write technical papers. Furthermore, as we can expect, TechCollo will be very useful for researchers doing interdisciplinary studies and having to check word combinations in an unfamiliar field of study.

We have made several plans for improving TechCollo. First, for pedagogical purposes, we plan to provide discipline-specific word lists on the TechCollo website. Those lists, compiled based on our domain-specific corpora, will be indexed with frequency information for various corpora (e.g. frequency in a medical corpus, in a cross-disciplinary academic corpus, or in BNC). We intend to embed our lists in an online interactive environment. Each entry (i.e. each listed word) is basically a link with which users can make further explorations on TechCollo. EAP students can conveniently click on a word and study its collocational patterns in different areas. The frequency information indexed to the words, moreover, will help users decide whether a word is particularly common and important in a domain or is rather frequent across several disciplines. Second, for technical purposes, we will continue to improve our techniques of extracting domain-specific collocations. For example, we plan to combine several association and statistics-based measures as Pecina (2008) did or use the ideas of paradigmatic modifiability by Wermter and Hahn (2005), and examine whether the revised techniques increase the precision of collocation extractions. We intend to investigate whether taking into account paradigmatic modifiability degrees and adding more association measures to TechCollo outperform the tradMI and normMI measures used by the current version of our tool. These new techniques will further be tested on various domain-specific corpora which may enable us to make some interesting discoveries in multiword terminology extraction.

REFERENCES

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*, 209-226.

Barrière, C. (2009). Finding domain specific collocations and concordances on the web. In I. Ilisei, V. Pekar, & S. Bernardini (Eds.), *Proceedings of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning* (pp. 1-8). Borovets, Bulgaria: INCOMA.

Campion, M. E., & Elley, W. B. (1971). *An Academic Vocabulary List*. Wellington: New Zealand Council for Educational Research.

Chen, Q., & Ge, G. C. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes*, *26*, 502-514.

Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. *System*, *32*, 251-263.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22-29.

Cohen, A., Glasman, H., Rosenbaum-Cohen, P. R., Ferrara, J., & Fine, J. (1979). Reading English for specialized purposes: Discourse analysis and the use of student informants. *TESOL Quarterly*, *13*, 551-564.

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

Coxhead, A. (2011). The Academic Word List 10 years on: Research and teaching implications. *TESOL Quarterly*, *45*, 355-362.

Davies, M. (2008). The Corpus of Contemporary American English (COCA): 400+ Million Words, 1990-present. Retrieved from: http://www.americancorpus.org

Farrell, P. (1990). *Vocabulary in ESP: A lexical analysis of the English of electronics and a study of semi-technical vocabulary* (CLCS Occasional Paper No. 25). Trinity Coll., Dublin (Ireland): Center for Language and Communication Studies.

Gardner, D., & Davies, M. (2013). A new Academic Vocabulary List. *Applied Linguistics*, *35*, 305-327.

Ghadessy, M. (1979). Frequency counts, word lists, and materials preparation: A new approach. *English Teaching Forum*, *17*(1), 24-27.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, *41*, 235-253.

Johansson, S. (1985). Word frequency and text type: Some observations based on the LOB corpus of British English texts. *Computers and the Humanities*, *19*(1), 23-36.

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* (Vol. 1, pp. 423-430). NJ, USA: Association for Computational Linguistics. doi: 10.3115/1075096.1075150

Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes*, *28*, 183-198.

Nagy, W., Anderson, R. C., Schommer, M., Scott, J. A., & Stallman, A. C. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly*, *24*, 262-282.

Nation, I. S. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Pecina, P. (2008). A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)* (pp. 54-57). Marrakech, Morocco: LREC.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, *19*(1): 143-177.

Vongpumivitch, V., Huang, J. Y., & Chang, Y. C. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes*, *28*(1), 33-41.

Wermter, J., & Hahn, U. (2004). Collocation extraction based on modifiability statistics. In *Proceedings of the 20th International Conference on Computational Linguistics* (pp. 980-986). Geneva, Switzerland: Association for Computational Linguistics.

Wermter, J., & Hahn, U. (2005). Paradigmatic modifiability statistics for the extraction of complex multi-word terms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 843-850). Vancouver: Association for Computational Linguistics.

West, M. P. (1953). *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London: Longman, Green.

Wible, D., Kuo, C. H., & Tsao, N. L. (2004). Improving the extraction of collocations with high frequency words. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)* (pp. 1855-1858). Lisbon, Portugal: European Language Resources Association.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, *3*(2), 215-229.

Yang, H. (1986). A new technique for identifying scientific/technical terms and describing science texts. *Literary and Linguistic Computing*, *1*, 93–103.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics* (pp. 189-196). Cambridge, MA: Association for Computational Linguistics.

CORRESPONDENCE

Ping-Yu Huang, General Education Center, Ming Chi University of Technology, Taiwan
E-mail address: alanhuang25@hotmail.com

Chien-Ming Chen, Institute of Information Science, Academia Sinica, Taiwan
E-mail address: virtualorz@gmail.com

Nai-Lung Tsao, Graduate Institute of Learning and Instruction, National Central University, Taiwan
E-mail address: beaktsao@stringnet.org

David Wible, Graduate Institute of Learning and Instruction, National Central University, Taiwan
E-mail address: wible@stringnet.org

PUBLISHING RECORD

Manuscript received: July 17, 2014; Manuscript accepted: October 17, 2014

發展及應用專業領域搭配詞搜尋學習工具

黃平宇
明志科技大學通識教育中心
陳建名
中央研究院資訊科學研究所
曹乃龍
國立中央大學學習與教學研究所
衛友賢
國立中央大學學習與教學研究所

Hyland & Tse (2007)指出，傳統的學術英語字彙表(例如：Coxhead，2000)忽略了一些重要的事實，即在不同的專業領域裡，學術字彙出現的比例不一，且常有不同的用法。舉例來說，在不同的學術領域中，學術字彙常與不同的字搭配出現，有時甚至呈現出不同的意義。因此，學術英語學習者需要的不是一套共用的學術英語字彙表，而是針對其學術領域特別整理的字彙知識。受到 Hyland & Tse 的啟發，我們發展了一項線上專業英語搭配詞搜尋工具，稱為 *TechCollo*。運用 TechCollo，學習者能夠比較在不同的專業英語語料庫裡，搭配詞出現的次數及使用的方式。此外，我們在此論文裡也使用 TechCollo 分析 Coxhead 的學術英語字彙在不同專業領域分佈的情形。我們的研究結果大致符合 Hyland & Tse 的論述，顯示學術英語字彙確實在不同專業領域出現的機率不一，且呈現明顯的搭配詞差異。

關鍵詞：專業領域搭配詞、專業領域語料庫、線上學習工具、學術英語