A Corpus-Based Tool for Exploring Domain-Specific Collocations in English

Ping-Yu Huang[1], Chien-Ming Chen[2], Nai-Lung Tsao[3] and David Wible[3]

[1]General Education Center, Ming Chi University of Technology
alanhuang25@hotmail.com

[2] Institute of Information Science, Academia Sinica
virtualorz@gmail.com

[3] Graduate Institute of Learning and Instruction, National Central University
{beaktsao, wible}@stringnet.org

Abstract

Coxhead's (2000) Academic Word List (AWL) has been frequently used in EAP classrooms and re-examined in light of various domain-specific corpora. Although well-received, the AWL has been criticized for ignoring the fact that words tend to show irregular distributions and be used in different ways across disciplines (Hyland and Tse, 2007). One such difference concerns collocations. Academic words (e.g. *analyze*) often co-occur with different words across domains and contain different meanings. What EAP students need is a "discipline-based lexical repertoire" (p.235). Inspired by Hyland and Tse, we develop an online corpus-based tool, *TechCollo*, which is meant for EAP students to explore collocations in one domain or compare collocations across disciplines. It runs on textual data from six specialized corpora and utilizes frequency, traditional mutual information, and normalized MI (Wible et al., 2004) as measures to decide whether co-occurring word pairs constitute collocations. In this article we describe the current released version of TechCollo and how to use it in EAP studies. Additionally, we discuss a pilot study in which we used TechCollo to investigate whether the AWL words take different collocates in different domain-specific corpora. This pilot basically confirmed Hyland and Tse and demonstrates that many AWL words show uneven distributions and collocational differences across domains.

1 Introduction

There has long been a shared belief among English for academic or specific purposes (EAP and ESP) instructors that it is necessary to provide students with a list of academic vocabulary [1] irrespective of their specialized domain(s). There are two main reasons why academic vocabulary receives so much attention in EAP instruction. First, academic vocabulary accounts for a substantial proportion of words in academic texts (Nation, 2001). Sutarsyah et al. (1994), for example, found that about 8.4% of the tokens in the Learned and Scientific sections of the Lancaster-Oslo/Bergen (Johansson, 1978) and Wellington corpora (Bauer, 1993). Second, academic words very often are non-salient in written texts and less likely to be emphasized by content teachers in class (Flowerdew, 1993). Consequently, EAP researchers have been convinced that students need a complete list of academic vocabulary, and several lists were thus compiled. Among the attempts to collect academic lexical items, Coxhead's (2000) Academic Word List (AWL) has been considered the most successful work to date. In the AWL, Coxhead offered 570 word families which were relatively frequent in a 3.5-million-token corpus of academic texts. The corpus was composed of writings from four disciplines: arts, commerce, law, and science. By considering certain selection principles such as frequency and range, Coxhead gathered a group of word families which were *specialized* in academic discourse and *generalized* across different fields of specialization. On average, the AWL accounted for 10% of Coxhead's academic corpus and showed distributions of 9.1-12% of the four disciplines. Since its publishment, the AWL has been frequently used in EAP classes,

[1] Academic words are also variously termed *sub-technical vocabulary* (Yang, 1986), *semi-technical vocabulary* (Farrell, 1990), or *specialized non-technical lexis* (Cohen et al., 1979) in the literature. They generally refer to words which are common in academic discourse but not so common in other types of texts.

Copyright 2013 by Ping-Yu Huang, Chien-Ming Chen, Nai-Lung Tsao, and David Wible
27th Pacific Asia Conference on Language, Information, and Computation pages 542－549

covered by numerous teaching materials, and re-examined by various domain-specific corpora (e.g. Vongpumivitch et al., 2009; Ward, 2009). The AWL, as Coxhead (2011) herself claims, indeed exerts much greater effects that the author ever imagined.

Although well-received, the AWL is not without criticisms. For instance, Chen and Ge (2007), while confirming the significant proportion of the AWL in medical texts (10.07%), found that only half of the AWL words were frequent in the field of medicine. In Hancioğlu et al. (2008), the authors criticized that the distinction that Coxhead (2000) made into academic and general service words was questionable. In actuality, there were several general service words contained in the AWL (e.g. *drama* and *injure*). Arguably the strongest criticism came from Hyland and Tse (2007), who questioned whether there was a single core academic word list. Hyland and Tse called Coxhead's corpus compilation "opportunistic" (p. 239) and built a new database better controlled for its selection of texts to examine Coxhead's findings. Utilizing a more rigorous standard, Hyland and Tse found that only 192 families in the AWL were frequent in their corpus. Furthermore, numerous most frequent AWL families did not show such high-frequency distributions in Hyland and Tse's dataset. In addition to these methodological problems, as Hyland and Tse emphasized, the AWL as well as those previous academic word lists ignored an important fact that words tend to behave semantically and phraseologically differently across disciplines. Many academic words, such as *analyze*, tend to co-occur with different words and contain different meanings across research areas. What EAP learners actually need and have to study, accordingly, should be "a more restricted, discipline-based lexical repertoire" (p. 235).

Inspired by Hyland and Tse's (2007) insights and analyses, we devise and create a learning tool which is able to generate domain-specific lexico-grammatical knowledge for EAP students. The knowledge that we focus on here concerns collocations. Specifically, we develop an online corpus-based tool, *TechCollo*, which can be used by EAP students to search for and explore frequent word combinations in their specialized area(s). The tool, by processing written texts in several medium-sized domain-specific corpora, enables students to study collocational patterns in their own domain, compare collocations in different disciplines, and check whether certain combinations or word usages are restricted to a specific field. To decide whether a pair of co-occurring words constitutes a candidate collocation, TechCollo uses measures such as frequency, traditional mutual information (MI) (Church and Hanks, 1990), and normalized MI (Wible et al., 2004). We will discuss these measures in more detail in Section 3.

This paper is structured as follows. In Section 2 we briefly discuss some related work. Section 3 describes the online learning tool and the corpora from which TechCollo extracts collocations. In Section 4, we present results of a pilot study to exemplify how to exploit TechCollo to discover differences in collocations across two domains. Finally, we propose our future plans for improving TechCollo in Section 5.

2 Related Work

In electronic lexicography or automatic term recognition (ATR), a number of studies have investigated how to retrieve multiword terminology from texts (e.g. Collier et al., 2002; Rindflesch et al., 1999). Basically, those studies identified candidate patterns of words (e.g. noun-noun or adjective-noun combinations) from texts and used various frequency-based or association-based measures to determine the *termhood* of those candidates. Other ATR studies took more sophisticated approaches. Wermter and Hahn (2005), for example, distinguished domain-specific from non-domain-specific multiword terms on the basis of *paradigmatic modifiability* degrees. The assumption behind this approach was that the component words of a multiword term had stronger association strength and thus any component of it was less likely to be substituted by other words. However, although the identification of multiword terms has been an active field of research, few studies have explored ways of making the terminology accessible to EAP students. To our knowledge, Barrière's (2009) TerminoWeb has been the only work addressing this issue in the literature. Below we describe Barrière's platform.

TerminoWeb, as its name suggests, was created with an aim to help learners of different professional areas explore and learn domain-specific knowledge from the Web. To get access to the knowledge, a user had to follow several steps. The starting point was to upload a technical paper to the platform. This paper was used as a source text in which the user selected

unknown terms and the TerminoWeb also automatically identified certain terms. Then, a set of queries were performed on the Web to collect texts relevant to the source text (i.e. belonging to the same domain) or including the same user-selected and computer-identified terms. Those collected texts were then a large domain-specific corpus. Within the corpus, the user could do concordance searches to understand word usages of an unknown term in larger contexts. The user could also make collocation searches for this 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 extracting collocations.

Unlike the technical corpora compiled via the TerminoWeb with texts from the whole Web and were likely to include lots of messy data, the corpora underlying TechCollo basically were composed of texts edited in advance which were assumed to be *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.

3 TechCollo: A Corpus-Based Domain-Specific Collocation Learning Tool

TechCollo, which stands for *technical collocations*, is an online tool with which EAP students can explore specialized collocations. To illustrate the functions of TechCollo, we respectively describe: (1) the compilation of ESP corpora underlying it, (2) the determination of a word pair as a candidate for a true collocation, and (3) the interface designed for EAP students.

3.1 Corpora

Currently, TechCollo extracts collocations from six domain-specific corpora. All of the six databases are medium-sized, containing 1.8-5.5 million running tokens. Among them, three were composed of texts coming from the largest online encyclopedia, Wikipedia. Specifically, the Wikipedia texts that we processed were provided by the Wacky team of linguists and information technology specialists (Baroni et al., 2009),[2] who compiled large Wikipedia corpora for various European languages such as English, Italian, and French. Based on an English corpus created by the Wacky team, we established corpora for three domains: medicine, engineering, and law, which were named Medical Wiki, Engineering Wiki, and Legal Wiki Corpora, respectively. The other three ESP textual archives contained writings from high-quality academic journals. That is, for the same medical, engineering, and legal domains, we consulted sixty academic journals and respectively downloaded 280, 408, and 106 articles from those journals online. We utilized the tools offered by Stanford CoreNLP (Klein and Manning, 2003) to POS-tag and parse the three academic corpora. The three corpora then were termed: Medical Academic, Engineering Academic, and Legal Academic Corpora.

In addition to the domain-specific corpora, TechCollo also provides collocation searches in two general-purpose corpora: Wikipedia and British National Corpus (2001). We offer collocation exploration for the two corpora for users to compare and identify collocations in subject areas and general use. Table 1 shows the corpus sizes of the six technical and two general-purpose corpora behind TechCollo.

Corpus	Token Count
Medical Wiki Corpus (MWC)	2,812,082
Engineering Wiki Corpus (EWC)	3,706,525
Legal Wiki Corpus (LWC)	5,556,661
Medical Academic Corpus (MAC)	1,821,254
Engineering Academic Corpus (EAC)	1,989,115
Legal Academic Corpus (LAC)	2,232,982
Wikipedia	833,666,975
British National Corpus (BNC)	94,956,136

Table 1: Sizes for Domain-Specific and General-Purpose Corpora

3.2 Collocation Extraction

In computational linguistics, various measures have been utilized in order to automatically extract collocations from texts. Those measures can be roughly divided into three categories (Wermter and Hahn, 2004): (1) frequency-based measures, (2) information-theoretical measures (e.g. mutual information), and (3) statistical

[2] The 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.

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, one is suggested to exploit several association measures with a correct understanding of their notions and behaviors.

TechCollo employs three main measures to decide whether a two-word combination constitutes a candidate collocation in a five-word window in our textual databases: frequency, traditional mutual information (*tradMI*) (Church and Hanks, 1990), and normalized MI (*normMI*, Wible et al., 2004). A learner using TechCollo can set or change the values of these measures to show candidate collocations in the six technical corpora (a detailed description of the user interface for TechCollo is given in section 3.3). First, the measure of frequency refers to raw co-occurrence count of a word pair. However, to filter out the pairs which are extremely frequent as a result of one or both of their component words but are not true collocations,[3] TechCollo offers the common association measure: *tradMI*, which is formulated as follows:

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

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, MI 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. introduces the alternative association measure *normMI*, which attempts to minimize the effects caused by sheer high frequency words. To achieve this, Wible et al. normalizes the *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 Wible et al. indicates, 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. compares the *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. It is important to note that, although the *normMI* produces higher recall than the *tradMI*, precision does not decrease accordingly. On our TechCollo interface, we provide the *normMI* to enable EAP learners to find and learn some word combinations which include high frequent words but are still true and specialized collocations in their domain(s).

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

3.3 User Interface

The main page of TechCollo is shown in Figure 1. Basically, this online collocation exploration tool allows users to choose from the six medium-sized domain-specific corpora: MWC, EWC, LWC, MAC, EAC, and LAC, and the two large-scale general-purpose corpora: BNC and Wikipedia. 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 (i.e. TechCollo 1.0) provides searches of verb-noun collocations. The

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

measures 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 MI or a longer list containing more candidate collocations.

Here we take the noun *procedure* and its verb collocates in MWC and EWC as examples. We feed this word into the TechCollo system with the frequency and *tradMI* set at 1 and 4, respectively. That is, only the verbs which appear together with *procedure* at least two times and having mutual information larger than 4 will be identified as candidate collocates. The search results are demonstrated in Figure 2.

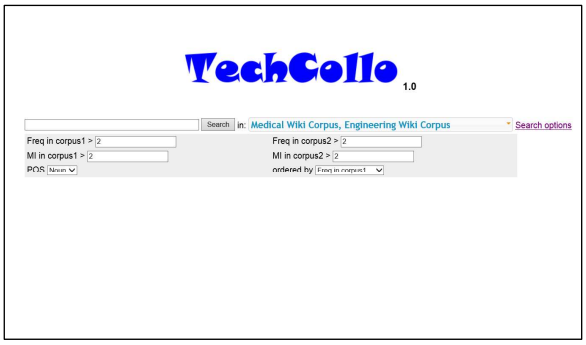

Figure 1: Main Page of TechCollo

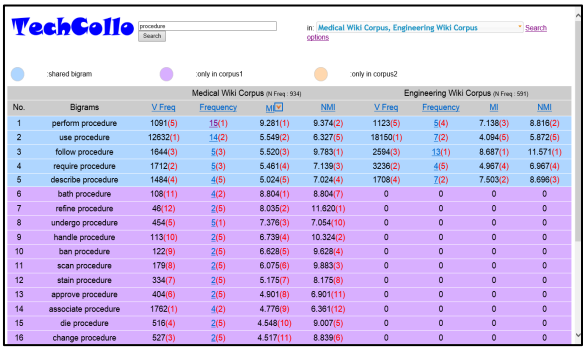

Figure 2: Search Results for *procedure*

According to the results offered by TechCollo, there are, respectively, 934 and 591 tokens of *procedure* in Medical Wiki and Engineering Wiki. Furthermore, the two corpora (or the two fields of profession) share several common collocations, including: *perform procedure*, *follow procedure*, *describe procedure*, etc. Taking a closer look at the *unshared* verb collocates in the two corpora (i.e. only in MWC or EWC), however, we find that *procedure* tends to co-occur with *undergo* and *die* only in MWC. These specialized collocations suggest that *procedure* is a technical term in medicine which refers to an operation. We expect and encourage EAP students to use TechCollo to explore and further discover such specialized collocations by: (1) searching collocations in a specific domain, (2) comparing collocations in two domain-specific corpora (e.g. MWC vs. EWC), and (3) comparing collocations in a specialized and a general-purpose corpora (e.g. MWC 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 offered by TechCollo 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 collocations containing high frequency component words. To examine the effectiveness of the *normMI*, we test it with certain legal collocations in the LAC, with the results shown in Table 2.

Collocation	*tradMI* ranking for the verb	*normMI* ranking for the verb
***break** law*	63	1
***push** trial*	14	7
***carry** obligation*	5	1

Table 2: Comparison of *tradMI* and *normMI* with Legal Collocations

In the three cases, specifically, we use the three nouns: *law*, *trial*, and *obligation* as keywords to search in the LAC and examine how the *tradMI* and *normMI* decide the rankings of the three high-frequency verb collocates: *break*, *push*, and *carry*. As Table 2 shows, *normMI* changes the rankings of these collocations with the three verbs being ranked in higher positions. The three verbs might not be noticed by learners using the *tradMI* and the *normMI* successfully raises them into more advantaged positions for learners. A more thorough examination, nevertheless, is required to investigate whether the *normMI* is indeed an effective measure of identifying collocations in domain-specific texts.

4. Comparing Collocational Patterns across Disciplines: A Pilot Study

To specify and illustrate how to use TechCollo in EAP studies, we ran a pilot study in 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 Academic Word List[4] (Coxhead, 2000) and explored and analyzed their verb collocates in the MWC and EWC. Our purpose, then, was to investigate whether it is true that words tend to show differences in collocations in different professional areas, as Hyland and Tse (2007) point 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 MWC and EWC. The very first observation that we made in the data generated by TechCollo was that many nouns showed uneven distributions in the two domain-specific corpora. Some examples of those nouns are given in Table 3. These distributional variations suggest that an academic word which is highly frequent and important in one discipline may be less important for students in another domain (e.g. the words *contractor*, *finance*, and *specification* for medical school students). EAP students who are required to study the AWL for their academic studies are very likely to be exposed to more lexical items than they actually need (Hyland and Tse, 2007).

Word	Frequency (per million tokens) in MWC	Frequency (per million tokens) in EWC
concept	115	332
contractor	1	53
contract	32	109
creation	35	90
datum	192	732
derivative	135	45
economy	18	100
evidence	329	93
finance	2	21
indication	104	29
methodology	13	60
policy	26	140
principle	96	214
processing	89	190
requirement	66	349
sector	10	135
specification	9	196
specificity	38	6
variable	25	128

Table 3: Nouns with Irregular Distributions in MWC and EWC

[4] As Coxhead (2000) explains, the word families of the AWL are categorized into ten sublists according to their frequency. Each of the sublists contains sixty families with the last one containing thirty.

In addition to the comparisons of numbers of occurrence, what interests us more concerns their relations with verbs in medicine and engineering. We present some of the verb-noun collocation data in Table 4.

Noun	Shared Collocates	Verbs in MWC Only	Verbs in EWC Only
analysis	perform		conduct
area		rub, scratch	
assessment			allow, perform
benefit	receive	confer	provide, offer
concept	use	employ	utilize
consistency		boil	
context	depend		
contract			negotiate, cancel
creation	result	induce	lead
environment	create		build
evidence	show	yield, reinforce	trace
factor		activate, inhibit	
formula		feed, determine	derive
function		affect, impair	replicate
issue	address	approach	deal
majority	make	constitute	
method	devise, employ		
policy			influence, implement
principle	operate		apply
procedure		undergo, die	
requirement	meet, fulfill		satisfy, comply
research	conduct	undergo	undertake
response	trigger, evoke	induce, stimulate	
role	play, fulfill		
structure	describe	elucidate, depict	
theory	develop, propose		formulate
variation	show	exhibit	display

Table 4: Verb Collocates in MWC and EWC

As Table 4 displays, there are several nouns which *share* verb collocates in the MWC and EWC, including: *context*, *method*, and *role*. In other words, these verb-noun combinations are of 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 even refer to different meanings across disciplines. The word *area*, for example, co-occurs with *rub* and *scratch* in MWC and not in EWC and refers to the specialized meaning of a part on the surface of human body. Several other nouns, such as *consistency, formula, function, procedure,* and *response* also contain such medicine-specific senses as they co-occur with *boil, feed, impair, die*, and *induce*, respectively. Another notable cross-disciplinary difference based on these collocations is, while expressing a similar idea, people in medicine and engineering appear to prefer different verbs. Examples for this include: *confer/offer benefit, employ/utilize concept, induce/lead creation, approach/deal issue, undergo/undertake research, exhibit/display variation,* etc. These field-specific idiomatic and habitual usages do not suggest that they are the only expressions that people in medicine or engineering use. Rather, they provide evidence showing that people in different areas tend to select different word combinations which form "a variety of subject-specific literacies" (Hyland and Tse, 2007: p.247). What EAP students need to study, then, should be these common specialized collocations and usages which make their writings and speech *professional* in their own domain(s).

5. Conclusion

The pilot study reported in this article basically 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 the AWL to learn academic words, they are very likely to memorize more lexical items than they actually need for studies in their own domain. Plus they will not be familiar with the common collocations that their colleagues frequently use in speech or writing. What the students need, or more specifically, what EAP researchers are suggested to develop, should be discipline-based vocabulary and collocation lists. Accordingly, 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. The tool, with its ability to allow students to learn specialized collocations in a discipline, compare collocations across disciplines, and explore collocations in domain-specific and general-purpose corpora, is of great help for EAP students to check word usages as they write technical papers. Furthermore, as we can expect, TechCollo will be very useful for researchers doing interdisciplinary studies and having to check word combinations across disciplines.

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 domains (e.g. in MWC, academic corpora, or BNC). EAP students can conveniently click on each listed word and study its collocational patterns in different areas. Second, for technical purposes, we will continue to improve our techniques of extracting domain-specific collocations. We plan to use the techniques and methods developed by, for example, Wermter and Hahn (2005) and Pecina (2008) and examine whether the revised techniques increase the precision of collocation extractions. Specifically, we intend to investigate whether taking into account paradigmatic modifiability degrees and combining several association measures outperform the *tradMI* and *normMI* measures used by the current version of TechCollo. These new techniques will further be tested on various domain-specific corpora which may enable us to make some interesting discoveries in terminology extraction.

Acknowledgements

The research reported in this paper was supported in part by a grant from Taiwan 's National Science Council, Grant #NSC 100-2511-S-008-005-MY3.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-crawled Corpora. Language Resources and Evaluation 43(3): 209-226.

Caroline Barrière. 2009. Finding Domain Specific Collocations and Concordances on the Web. Proceedings of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning.

Laurie Bauer. 1993. Manual of Information to Accompany the Wellington Corpus of Written New Zealand English. Victoria University of Wellington.

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

Qi Chen and Guang-Chun Ge. 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(4): 502-514.

Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics 16(1): 22-29.

Andrew Cohen, Hilary Glasman, Phyllis R. Rosenbaum-Cohen, Jonathan Ferrara, and Jonathan Fine. 1979 Reading English for Specialized Purposes: Discourse Analysis and the Use of Student Informants. TESOL Quarterly, 34: 551-564.

Nigel Collier, Chikashi Nobata, and Junichi Tsujii. 2002. Automatic Acquisition and Classification of Terminology Using a Tagged Corpus in the Molecular Biology Domain. Terminology 7(2): 239-257.

Averil Coxhead. 2000. A New Academic Word List. TESOL Quarterly, 34(2): 213-238.

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

Paul Farrell. 1990. Vocabulary in ESP: A Lexical Analysis of the English of Electronics and a Study of Semi-Technical Vocabulary. CLCS Occasional Paper No. 25.

John Flowerdew. 1993. Concordancing as a Tool in Course Design. System 21(2): 231-244.

Nilgün Hancioğlu, Steven Neufeld, and John Eldridge. 2008. Through the Looking Glass and into the Land of Lexico-Grammar. English for Specific Purposes 27(4): 459-479.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1.

Ken Hyland and Polly Tse. 2007. Is there an "academic vocabulary"?." TESOL Quarterly 41(2): 235-253.

I. S. P. Nation. 2001. Learning Vocabulary in Another Language. Cambridge University Press, Cambridge, UK.

Stig Johansson. 1978. Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers. University of Oslo. Oslo, Norway.

Pavel Pecina. 2008. A Machine Learning Approach to Multiword Expression Extraction. Proceedings of the LREC MWE 2008 Workshop.

Thomas C. Rindflesch, Lawrence Hunter, and Alan R. Aronson. 1999, Mining Molecular Binding Terminology from Biomedical Text. Proceedings of the AMIA Symposium. American Medical Informatics Association.

Frank Smadja. 1993. Retrieving Collocations from Text: Xtract. Computational Linguistics 19(1): 143-177.

Cucu Sutarsyah, Paul Nation, and Graeme Kennedy. 1994. How Useful Is EAP Vocabulary for ESP? A Corpus Based Case Study. RELC Journal 25(2): 34-50.

Viphavee Vongpumivitch, Ju-yu Huang, and Yu-Chia Chang. 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.

Jeremy Ward. 2009. A Basic Engineering English Word List for Less Proficient Foundation Engineering Undergraduates. English for Specific Purposes 28(3): 170-182.

Joachim Wermter and Udo Hahn. 2004. Collocation Extraction Based on Modifiability Statistics. Proceedings of the 20th International Conference on Computational Linguistics. Association for Computational Linguistics.

Joachim Wermter and Udo Hahn. 2005. Paradigmatic Modifiability Statistics for the Extraction of Complex Multi-word Terms. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

David Wible, Chin-Hwa Kuo, and Nai-Lung Tsao. 2004. Improving the Extraction of Collocations with High Frequency Words. Proceedings of International Conference on LREC.

Huizhong Yang. 1986. A New Technique for Identifying Scientific/Technical Terms and Describing Science Texts. Literary and Linguistic Computing 1: 93-103.

David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics.