ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 11, No. 3, pp. 480-485, May 2020
DOI: http://dx.doi.org/10.17507/jltr.1103.17

The Study and Application of Corpus Linguistics to Create Content-based Materials for Pedagogical Purposes

Zhenyu Yang
Foreign Languages College, Inner Mongolia University, Hohhot, China

***Abstract*—It has long been proved that content-based instruction affects the overall academic progress of ESL students. With the development of research and application of corpus linguistics, such materials can be better collected and sorted for pedagogical purposes. In this project, a creation of a corpus of business language was used for developing materials for the ESL class. After a process of piloting, experimenting and revising, the materials were developed based on the collected corpus. As a result, the course turned out to be effective and well received by both the instructor and the students. We believe this project could be replicated for other academic disciplines as well.**

***Index Terms*—corpus linguistics, content-based material, ESL**

I. Introduction

Every year, large numbers of international students take part in language training programs in the university in the U.S. to prepare them for their graduate studies. Even though the program focuses on English for academic purposes (EAP), the transition to discipline-specific coursework can prove to be problematic when students are confronted with the English for specific purposes (ESP) of the various disciplines. Many of the international students who come to the U.S. or other major English-speaking countries to pursue academic careers choose Business Administration as a major. These students are required to take and pass the introductory course in the School of Business Administration in order to continue with business classes. In the university I did research at, there is a high failure rate of about 30% every year for the international students who enroll, which constitute a major concern for both the teachers at the language training program and those at the business administration. As a collaborative attempt to solve the problem, I participated in a project to create new teaching materials and methods at the university in the U.S. The suggestion was to utilize a corpus as a useful resource for developing classroom materials. Such a computer-based, principled collection of the actual language encountered in the class could provide insight into the language that the students would need to master.

II. Literature Review

A. Content-based Classes

Content-based classes refer to those that teach language skills through the lens of the particular academic course or discipline with which they are paired. Kasper (1994) sums up the situation commonly faced by ESL students:

> Given the sophistication and complexity of the ideas and material presented in…academic courses, and the fact that many of them are taught lecture style, a great number of ESL students find themselves overwhelmed and frustrated in these classes. They are expected to perform at the level of native speakers even though their proficiency in English is often inadequate to this task.
>
> To increase their chances for academic success, they need a situation that will provide them with additional practice with the ideas presented in their content classes while helping them to improve their English language skills. Pairing the ESL class with the academic content class offers just such a situation. (p. 376)

Kasper has made numerous investigations into the relationship between content-based ESL classes and student performance. The 1994 research found that students in these paired classes did well in the academic class with which their ESL class was paired: 78% received a B or better, and none failed. Additionally, students enrolled in ESL classes paired with regular academic courses scored significantly higher in their English language reading skills than did students in non-paired ESL classes. Similar results would be welcomed by both the ESL programs and the School of Business Administration

In further research, Kasper (1997) investigated how content-based instruction affected the overall academic progress of ESL students in order to confirm student feedback suggesting that those who had taken content-based ESL classes had not only graduated at a higher rate but also done so with better grades. This multi-term study followed the performance of students enrolled in content-based ESL classes until they graduated and compared them with ESL students who had not been in content-based classes. In the four terms of the study, students in the content- based group scored higher on several measures: they were found to have done significantly better in the ESL class, they scored significantly higher on a reading assessment examination, and they graduated at a higher rate than those in the control group who weren't in content-based classes.

B. *Corpora in Content-based Language Study*

If content-based instruction is indeed successful in improving the performance of ESL students, the question naturally arises as to how the content is to be selected for a given paired course. With the development of research and application of corpus linguistics, a new answer has been given: Biber and Barbieri (2007) argue that the language that students need to control is that which they will encounter in their studies. Conrad (1996) and Sznajder (2010), too, advise that the language taught be that which is found in authentic texts from the disciplines.

Several instances of using corpora of business English for pedagogical purposes have been described in the literature. Fuentes (2002) utilized a business corpus of more than a million words to inform his creation of course materials based on the frequencies and collocates of content words he found, after which he evaluated the materials in order to revise and improve what he had done. His conclusion was that "texts recommended or required at the university should…serve as reference…" for the creation of materials (p. 28), a conclusion similar to that of Conrad and Sznajder.

In a case study of adult learners in Germany, Walker (2011) found that having students work directly with corpora informed the teaching of word choice in business settings, specifically for showing shades of meanings of collocates of business-related words used in specific situations anticipated by the students themselves. Liu (2010) searched the Corpus of Contemporary American English (COCA) for near-synonyms of words before describing a behavioral profile for each one while Diani (2008) searched for a specific word and then analyzed its use, focusing on its collocates. Although these latter studies were not all limited to business language, they did suggest additional ways of using corpora to inform pedagogy.

When multiple corpora are used in research, or when large corpora are divided into subcorpora, it is possible to investigate differences in the language among the various corpora or subcorpora. One of the most extensive investigations of this type was done by Coxhead (2000), in developing an Academic Word List (AWL) of 570 word families by comparing word frequencies in an academic corpus built by the researcher to those in a general service list. The AWL that was created represented 10% of the words in the academic corpora but only 1.4% in a more general corpus, indicating their increased salience in academic English. While Coxhead herself didn't report using her results for pedagogical purposes, Schmitt and Schmitt (2005) employed this AWL to create a workbook for mastering academic vocabulary.

Thus far we have seen that research on content-based classes supports the formulation of a class specifically for the ESL students who enroll in the introductory business class and that corpus research suggests ways of identifying language for the creation of materials drawn from authentic materials. Identifying vocabulary used in the introductory business class could well inform the development of materials for the ESL class. Comparing the language of the textbook with that of online business blogs or with other genres could provide valuable input for additional pedagogical materials. Familiarizing ESL students with multi-word strings that are used frequently in business could improve their writing.

III. Data Collection

The design of the corpus was undertaken in the spring, which underwent both peer and instructor review and subsequent revision. Several of the graduate students who had worked on the design of the corpus committed to building it. The group decided to concentrate on entering the textbook as it was used in all of the sections of the course, was the basis of tests and quizzes, and most of the reading in each section was assigned in it.

During the spring and summer of the same year samples from the textbook, newspaper articles and blogs were entered into the corpus. Most of the data were in the electronic form, which only need to be cleaned and stored in a logical order. The result was a creation of a corpus of business language to be used for developing materials for the class. The corpus as it stands has samples from three genres: the textbook, business articles from the online version of some major newspapers, and business blogs. Samples of approximately five hundred words were taken from each section of each chapter of the textbook and included any titles for these sections. Since sentences were not cut off at 500 words and some sections were shorter than the planned size, samples varied in actual size. Chapter summaries were entered in their entirety, and graphics and charts were not included. The newspaper articles and blogs were collected whole. The components and size of the corpus are summarized in Table I.

Table I.
Contents Of The Corpus As Of August 2019

Text Type	Number of Samples	Number of Words
Textbook	442	107,979
Business Articles	48	31,181
Business Blogs	157	56,888
Total	625	196,048

IV. Methods

The project had two phases: in the first phase, target vocabulary was identified in the corpus and in the second phase materials were created using this target vocabulary. The goal was to help bridge the gap between the academic vocabulary the ESL students knew and the discipline specific vocabulary of the course.

A. Phase I. Identifying Target Words and Deciding When to Introduce Them

The first phase of the project was divided into two parts: the first was selecting the word families to be introduced as target vocabulary, and the second was determining when to introduce them. During the first part, we searched the textbook subcorpus as a whole for word frequencies using the MonoConc Pro concordancer (Barlow, 2003). The initial reports run were simple frequencies. However, the reports showed 980 tokens (individual occurrences) of the word *firms*, making it the 12th most frequent word in the corpus, and 854 tokens of *firm*, making it the 15th most frequent word. In general, the most frequent words in corpora are function words; it is unusual to find content words this common in a corpus. (See Table II below)

TABLE II
RESULTS OF SEARCH FOR SIMPLE FREQUENCIES SHOWING THE 15 MOST COMMON WORDS IN THE CORPUS

Number of Tokens	Percentage in corpus	Word
6417	5.9428%	the
3596	3.3303%	to
3214	2.9765%	of
2946	2.7283%	a
1938	1.7948%	and
1744	1.6151%	in
1586	1.4688%	that
1385	1.2827%	or
1315	1.2178%	is
1181	1.0937%	are
1035	0.9585%	be
982	0.9094%	**firms**
922	0.8539%	may
856	0.7927%	as
860	0.7965%	**firm**

This changed the focus of the searches from individual words to word families. Function words were not considered for inclusion in the vocabulary list. We identified approximately 80 high frequency word families as candidates for inclusion in the vocabulary list then took them to the instructors for their comments. In our meeting, several of the families in the initial list were eliminated because the instructors felt that the students would already be familiar with them. Considering the duration of the course for the quarter and the acceptance of the students, the list was reviewed again and culled down to fit the constraints of the class. As a result, about 60 word families were included in the list.

The second part was to determine when to introduce each word family, that is, to assign the target vocabulary to the six lessons. The main difficulty encountered with this was that most of the target vocabulary appeared throughout the entire textbook, beginning with the first chapter. Introducing words the first time they appeared was, therefore, not a reasonable option: it would mean that the first lesson would have most of the vocabulary in it, exceeding our limit on the number of word families in a single lesson. To solve this problem of when to introduce the families, we had a pilot study and discussion with the instructors, and finally decided to group the target vocabulary based on its frequencies in the six parts into which the textbook was divided: Business Environment, Starting a New Business, Management, Managing Employees, Marketing, and Financial Management. This would align the presentation of the word families with the order of the material in the textbook. We developed a spreadsheet that showed their distribution throughout the textbook. This approach provided a workable solution to the distribution problem.

The spreadsheet listed the total tokens (occurrences) of the word families in the corpus and the percentages of the tokens found in each of the six parts of the textbook. Table III below shows the first six families (represented by a single family member) as an example. This section of the spreadsheet shows that 66% of the 96 tokens of the word family of “corporate” are in Part II of the textbook. This word family was, therefore, placed in Part II. Assignments of words like “corporate” were simple as were those of words like decentralize and franchise which appeared in only one part of the textbook, but other families were more widely disbursed. In general, those widespread families were placed in the part where their cumulative frequencies reached approximately 50%.

TABLE III.
SPREADSHEET SHOWING A SAMPLE OF THE DISTRIBUTION OF VOCABULARY WORDS THROUGHOUT THE TEXTBOOK (SHADED BACKGROUND INDICATES WHERE FAMILY IS INTRODUCED)

	consume	**corporate**	**credit**	**customer**	**decentralize**	**distribute**
Total tokens	**254**	**96**	**166**	**409**	**35**	**178**
Tokens / 100,000	**235**	**89**	**154**	**379**	**32**	**164**
% in						
Part I	**38**	**14**	**23**	**26**	**0**	**6**
Part II	**2**	**66**	**12**	**14**	**0**	**7**
Part III	**1**	**9**	**2**	**13**	**100**	**2**
Part IV	**0**	**3**	**1**	**7**	**0**	**0**
Part V	**59**	**2**	**20**	**37**	**0**	**77**
Part VI	**0**	**6**	**43**	**3**	**0**	**8**

B. Phase II. Developing and Revising Materials

After identification and assignment of the target words, we came to the second phase: developing and revising materials. Several principles were followed for the presentation of the target vocabulary:

1. authentic materials from corpora would be used throughout
2. target vocabulary would be presented multiple times
3. materials would appear in consistent formats for each lesson

The goal in following these principles was to flood the students with the vocabulary that they would be expected to learn in the course. Using a consistent format for each lesson would allow the students to focus on the content of what was being presented rather than its format. This approach was consistent with that taken in the two books used as models for the materials: Robbins' (2006) Co*llins COBUILD Business Vocabulary in Practice* and Schmitt & Schmitt's (2005) *Focus on Vocabulary: Mastering the Academic Word List*, both of which are corpus based.

The overall process for developing and assessing the materials lasted over three quarters. The first step was to try out a number of different formats for exercises in the spring quarter, getting feedback from the instructor on what worked best. Based on the feedback, a more complete set of materials was prepared and piloted in the class fall quarter, when feedback from the students would be solicited. Revisions of materials were done and a final set prepared for the winter quarter.

The two examples of usage of each word given were taken from the Corpus whenever possible. When the Corpus did not offer an example, one was found in the Corpus of Contemporary American English (COCA), which is a free, online corpus of over 450 million words. Examples drawn from COCA were limited to those which showed non-business-specific meanings of words. The definitions given were not limited to business uses of the target vocabulary as both of the instructors and I were committed to introducing the students to meanings outside of a strict business context so students would have some understanding of broader uses of these words. An example of this is the use of *firm* meaning "strong" as in "He offered the most annoying form of handshake – dismissive, minus the ***firm*** grip or eye contact". Here the student needs to understand that a "firm grip" is not some secretive way of shaking hands prescribed by a company.

All of the exercises were constructed using only sentences from the Corpus. Again, the reason for this was to repeat information from the textbook and to display target words in context. Sentences chosen for the exercises were always different from those in the examples of vocabulary usage.

V. RESULTS

Graduate students in the MA TESOL program have been working actively on the Corpus project. After three quarters of piloting, experimenting and revising, the percentages of the students passing the exam, either the subjective or objective tests, were both increased. (From 72% and 76% to 77% and 80% respectively) When the students did the matching exercise of the vocabulary review for the final the results showed that they knew most of the vocabulary well. We also interviewed some of the students at the end of the different quarters for their feedback, and most of them said the new materials help them memorize the vocabulary better. The content-based materials turned out to be effective and well received by both the instructor and the students.

VI. DISCUSSIONS

Too often we are assigned materials to use that are more prescriptive than descriptive and are not based on actual usage, when it is often the later our students need to understand. Too many sample sentences and dialogs appear to be manufactured rather than authentic and don't prepare students for what they will actually hear. I have come to appreciate and embrace the necessity of having students understand how people really speak and write, and this is what is found in corpora.

Generally, the time spent with this Corpus project has been incredibly valuable beginning with designing what would be included in the corpus; through countless hours of scanning the textbook for inclusion in the corpus, counting words, looking up words, and finding examples of word use; and puzzling about how to distribute the vocabulary. We are pleased that the results of this project are, indeed, useful for the students for whom they were written and we sincerely hope that the materials will benefit those who will take this course in the coming years and the instructors and students taking similar ESL courses in similar contexts.

My personal assessment of this project is that it has two major strengths. The first is that a solid foundation has been established upon which other materials can be built for business language for the program, whether for another iteration of the paired class, the elective business class that is currently offered, or in other classes when business is the topic being discussed. There now exists a large amount of readily accessible information to which instructors can go to create business-related materials.

On a grander scale, the project also established a process for developing vocabulary materials for other similar classes in the future. Business language course is not the only one that other international students have difficulty with when they take regular academic classes. Given the opportunity to do the materials again, I think I would repeat what was done, but would include more work on collocations, information on common multi-word strings, and additional lessons teaching students how to use corpora. While the information and materials that were created could prove useful for business related classes, corpus skills can be used by students on their own in order to answer a broad range of questions about language.

On a larger scale, this project could be replicated for other academic disciplines such as psychology, medicine, engineering, or sociology. If international students have difficulties in the introductory business course, it would not be surprising to find that they have difficulty in other introductory courses as well.

VII. Limitations

The corpus itself would become even more useful were it to be expanded to include other subcorpora as outlined in the original design. These might include articles from other business publications, transcripts of lectures, etc., many of which are readily available online and could be collected with a minimum of time spent. Comparing the language in these additional subcorpora to that of the textbook likely uncover differences between the genres.

A broad research project that could be done with paired content-based classes that uses corpus-based materials would be to follow ESL students from the class throughout their university careers to see how they perform compared to ESL students who have not had such classes. This would essentially replicate the work done by Kasper. If there is evidence that the ESL students who have this type of class perform better than those who don’t, it would argue for the inclusion of more classes like this in our program.

References

[1] Barlow, M. (2003). Concordancing and corpus analysis using MP 2.2. Houston, TX: Athelstan.

[2] Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics, 14*, 275-311.

[3] Biber, D. & Barbieri, F. (2007) Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*, 263-286.

[4] Conrad, S. (1996). Investigating academic texts with corpus-based techniques: An example from biology. *Linguistics and Education, 8*, 299-326.

[5] Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*, 213 – 237.

[6] Davies, M. (2008-) The corpus of contemporary American English (COCA): 410_ million words, 1990-present. Retrieved from http://www.americancorpus.org (accessed 10/11/2019).

[7] Diani, G. (2008). Emphasizers in spoken and written academic discourse: The case of *really*. *International Journal of Corpus Linguistics, 13*, 296-321.

[8] Fuentes, A. C. (2002). Exploitation and assessment of a business English corpus through language learning tasks. *ICAME Journal*, 26, 5-30.

[9] James, M. A. (2006). Transfer of learning from a university content-based EAP course. *TESOL Quarterly*, 40, p. 783-806.

[10] Kasper, L. F. (1994). Improved reading performance for ESL students through academic course pairing. *Journal of Reading*, 37, 376-384.

[11] Kasper, L. F. (1997). The Impact of content-based instructional programs on the academic progress of ESL students. *English for Specific Purposes, 16,* 309-320.

[12] Kasper, L. F. (1995/6). Using discipline-based texts to boost college ESL reading instruction. *Journal of Adolescent & Adult Literacy, 39*, 298-306.

[13] Kozlov, M. & Engelmann, T. (2015) Is knowledge best shared or given to individuals? Expanding the Content-based Knowledge Awareness Paradigm. *Computers in Human Behavior*, *51*, 15-23.

[14] Kramer, J. (2011). Creating the BA 101 corpus and a guide for its use. (Unpublished MA TESOL project report), Portland State University, Portland, OR.

[15] Lemay, C., L’ Homme, M., & Drouin, P. (2005). Two methods for extracting “specific” single-word terms from specialized corpora: Experimentation and evaluation. *International Journal of Corpus Linguistics, 10,* 227-255.

[16] Liu, D. (2010). Is it a chief, main, major, primary, or principal concern? A corpus-based behavioral profile study of the near-synonyms. *International Journal of Corpus Linguistics, 15*, 56-87.

[17] Madura, J. (2010). Introduction to business. St. Paul, MN: Paradigm Publishing, Inc.

[18] Peake, B. (2011). Formulaic language for academic writing: analysis and materials development with text organizing lexical bundles. (Unpublished MA TESOL project report). Portland State University, Portland, OR.

[19] Robbins, S. (2006). Collins COBUILD business vocabulary in practice. Glasgow, Great Britain: HarperCollins Publishers, Ltd.

[20] Schmitt, D. & Schmitt, N. (2005). Focus on vocabulary: Mastering the academic word list. White Plains, NY: Pearson Education, Inc.

[21] Simpson, R. & Mendis, D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly, 37*, 419-441.

[22] Sznajder: H. S. (2010). A corpus-based evaluation of metaphors in a business English textbook. *English for Specific Purposes*: *29*, 30-42.

[23] Walker, C. (2011). How a corpus-based study of the factors which influence collocation can help in the teaching of business English. *English for Specific Purposes, 30*, 101-112.

Zhenyu Yang is an associate professor at Foreign Languages Department of Inner Mongolia University. His main research interests are computer-assisted language learning, TESOL and corpus linguistics. He did research in the United States as a visiting scholar for one year.