Scholar Publishing Group
International Journal of Educational Curriculum Management and Research
https://doi.org/10.38007/IJECMR.2023.040210
ISSN 2790-0967 Vol. 4, Issue 2: 82-90

# *A Study of Corpus-based Academic Formulas and Its Postgraduate Teaching*

**Dongmei Meng[1*] and Chunming Meng[2]**

[1]*School of Foreign Languages, East China Jiaotong University / Research Center for Applied Translation of Transportation and Engineering, Nanchang, Jiangxi 330013, China*

[2]*Jiangxi College of Foreign Studies, Nanchang, Jiangxi 330013, China*

*605840009@qq.com*

[*]*corresponding author*

***Keywords:*** Academic Formula, Corpus-Based, Postgraduate Teaching

***Abstract:*** With the rapid development of modern educational technology, corpus with the concept and technical features of big data has played an important role in the field of language teaching and research. The study investigates the use of corpus resources in teaching English program language and introduces the concepts and function of academic formula, outlines the research on language teaching using corpus, analyzes the types of corpus, and gives an overview of the common corpus resources for teaching English in college. The study outlines the research on language teaching using corpus, analyzes the types of corpus and gives the common corpus resources for teaching English in university.

## 1. Introduction

With the rapid development of modern educational technology, corpus with the concept and technical features of big data has played an important role in the field of language teaching and research, and corpus linguistics has grown into an emerging school of interdisciplinary interaction between modern education, linguistics and computer science. With the characteristics of large corpus size, authentic sources, representativeness, and easy access and analysis, corpus has become an indispensable resource and tool for language research and language teaching. Therefore, the scientific and reasonable use of corpus resources in academic English teaching is both an inevitable trend of modern educational technology development and an endogenous demand for academic English teaching reform.

As early as the 1980s, some experts and scholars proposed foreign language teaching based on corpus linguistics, believing that foreign language teaching should start from the most frequently occurring words and the most commonly used collocations in the target language. In 1997, a symposium on "Corpora and Translation Learning" held in Italy focused on the importance of using monolingual corpora, analogical corpora, and parallel corpora in translator training.

Chinese corpus research began in the late twentieth century. In terms of English corpus

construction, the JDEST (Science and Technology English Corpus of Shanghai JiaoTong University), which Professor Yang Huizhong presided over in the 1980s, was the earliest English corpus in China. Due to the increasing popularity of corpus in English teaching, scholars' interest in this research has been increasing in recent years. Currently, many scholars in China have different research directions on the corpus-based textbook vocabulary research, such as focusing on the comparison and analysis of collocation information in English vocabulary teaching; the distribution and usage characteristics of vocabulary in textbooks as a whole; and the ways and methods used to organize English specialized vocabulary teaching, etc. In 2004, some scholars used the methods of corpus index co-occurrence dynamic context, text equivalence probability analysis, translation style quantification and multi-translation In 2004, a scholar constructed the idea of corpus translation teaching through the methods of corpus index co-occurrence dynamic context, text equivalence probability analysis, translation style quantification and multiple translations. In the same year, some scholars demonstrated the application value of parallel corpus in translation teaching, and in 2008, some scholars made a preliminary analysis of the current situation of research on English translation materials and translation teaching based on corpus. It should be noted that the above-mentioned studies mainly focus on translation and vocabulary teaching, and most of them are based on self-built small corpora, which have certain limitations in terms of authority, corpus richness and representativeness, so it is difficult to achieve the expected teaching effect. This study explores the extraction and teaching of English academic program language based on corpus from the concept and function of academic program language, and provides reference for academic English teachers.

## 2. Definition, Function and Extraction of Academic Formula

### 2.1. Definition of Academic Formula

In the English literature, there are many terms for the concept of "academic formulas", and different researchers have given it different names. Among them, the common names are: academic formulas, idiomatic chunks, lexical bundles etc[6]. The types are often classified from the perspective of usage and acquisition: according to usage, they can be classified as greetings, good-byes, elaborations, instructions, expressions, declarations, exclamations, inquiries, and proverbs[11]; according to second language acquisition, they can be classified as non-analytic, partially analyzable, and analyzable chunks. Wary[14] first introduced the concept of academic formulas: "a continuous or discontinuous sequence of words or other meaningful components that is or appears to be prefabricated, i.e., when used is extracted or stored from memory as a whole rather than generated or analyzed through the grammar of the language." This is one of the most authoritative and wide-ranging definitions available. Based on the previous research, our scholar Lu[7] believes that a academic formula is "a continuous or discontinuous multi-word unit or framework that is understood and produced by adult native speakers in a holistic form, which occurs more frequently in language use and performs certain syntactic and/or pragmatic functions". In the author's opinion, academic formula is a fixed expression that can improve the quality of learners' language output, including a variety of forms such as proverbs, fixed sentence patterns, and stable collocations. For high school learners, a grapheme is "an idiomatic phrase that is directly extracted from the native language and can be directly applied to oral or written language output without grammatical analysis."

### 2.2. Functions of Academic Formula

The most representative studies in functional research include Wray & Perkins[4] discusses the

four functions of academic formula: firstly, academic formula is a tool for social communication; Secondly, academic formula is a shortcut for discourse processing; Thirdly, academic formula is beneficial for the generation and understanding of discourse; The fourth type is academic formula, which is a symbol of children's growth. Handle & Graf[5] claims that the functions of academic formula are reflected in the following three aspects: firstly, from a cognitive perspective, academic formula can reduce the cognitive load of the speaker; Secondly, from a pragmatic perspective, academic formula constitutes a part of native language communicative competence; Thirdly, from a developmental perspective, academic formula is an important 'acquisition assistant'. It is widely believed in the academic community that children are indeed able to use complex word strings or academic formulas in the early process of acquiring their mother tongue. Wray[14] divided the academic formula that appears in children's language into two categories: one is inanalyzed strings, and the other is fused strings. Unanalyzable word strings refer to word strings that repeatedly appear in a fixed form and perform specific functions in the same context. These word strings are remembered and used by children as a whole, but their meanings may not be understood by children. Fusible word strings refer to word strings that are constructed, stored, and used as a whole by children based on their already acquired knowledge of grammar and vocabulary. In the process of children's growth, these two academic formulas alternate to meet the needs of children to express their own needs and win the recognition of caregivers and the community they belong to; On the other hand, it can reduce the burden of language processing on children, thereby enhancing their fluency in language use. Perera[10] found through tracking the use of academic formula in the learning process of four Japanese children that academic formula plays an indispensable role in the initial stage of second language acquisition. It is the foundation for innovative language use, and together with innovative language, it forms a living and functional language.

### 2.3. Extraction of Academic Formula

Durrant[2] and Ackermann & Chen[1] have successively compiled lists of common collocations in general academic English based on different corpora. These three collocation lists ended up with differences in their respective contents due to their focus on different target collocations and different collocation screening methods(Table 1).

*Table 1. Extraction of academic formula*

| Selection Method | researcher | Extraction form |
|---|---|---|
| Matching type | Durrant | Mostly a grammatical collocation of a function word plus a real word |
| | Ackermann & Chen | Limited to lexical collocations, i.e. adjectives with nouns or nouns with nouns, verbs with nouns or verbs with adjectives, verbs with adverbs, etc. |
| Language sources | Durrant | Academic papers in five subject areas |
| | Ackermann & Chen | Academic papers and textbook chapters in 28 disciplines |
| | Lei & Liu | Mainly journal articles, book reviews, and theses of native English-speaking PhD students in applied linguistics at British and American universities. |
| Identification and extraction criteria for collocation | Durrant | Non-academic corpus from the BNC corpus as a comparison benchmark |
| | Ackermann & Chen | Frequency and mutual information values were also used in screening the pairings, with additional dispersion rates and higher t-value parameters ( t $\geqslant$ 4) |

In terms of collocation types, most of the collocations in Durrant's[2] list are grammatical collocations of a function word plus a real word, such as this study, based on, etc. Ackermann & Chen's[1] list is limited to lexical collocations, i.e., adjectives with nouns or nouns with nouns, verbs with nouns or verbs with adjectives, verbs with adverbs, etc. Lei & Liu's list[6-7] is also a lexical collocation list like Ackermann & Chen's[1], but they deliberately ensure a balanced number of collocation types. In terms of corpus sources, Durrant's[2] corpus consists of academic papers in five subject areas, Ackermann & Chen's[1] corpus consists of academic papers and textbook chapters in 28 subjects, while Lei & Liu's[6-7] corpus mainly consists of journal articles, book reviews, and theses of native English-speaking PhD students in applied linguistics at British and American universities. Thesis. In terms of collocation identification and extraction criteria, the commonly used corpus parameters include frequency, mutual information value (MI), t value, dispersion, keyness, etc. Durrant's[2] study used the non-academic corpus in the BNC corpus as a benchmark for comparison, and by keyness The study by Durrant[2] used the non-academic corpus of the BNC as a benchmark for comparison, and selected 1,000 key collocations from the academic corpus, as well as those with high frequency and high mutual information value in five academic disciplines. As in Durrant's[2] study, Ackermann & Chen[1] also used frequency and mutual information values in screening collocations, and added additional parameters of dispersion rate and higher t-value ($t \geqslant 4$) because they found that collocations below the set t-value criterion were usually non-target collocations of noun-preposition collocations or fragments of longer phrases[3].

## 3. Postgraduate Teaching Based on Corpus

Learners of academic English often choose appropriate language forms (including academic formula) according to the meaning and function they need to convey in the process of academic writing or oral expression. Therefore, the top-down, function-first approach is generally adopted in corpus-based academic English research, especially in the field of genre analysis, in which genre structure is described and analyzed through manual annotation of steps and steps, and the analysis of linguistic form is given secondary importance. Both form-first and function-first paradigms have produced many valuable results for academic English language teaching, but a single focus on form or function can lead to a "function-form" gap[8]. Lu Xiaofei[7] argues that this approach should also be applied to academic formula teaching and research in order to reveal the "form-function" mapping of academic formula and help learners to learn English. This approach, according to Lu Xiaofei[7], should also be applied to the teaching and research of academic formula in order to reveal the "form-function" mapping of academic formula and help learners acquire the knowledge to use it effectively in academic English discourse. The authors argue for the use of genre analysis, research domain analysis, and corpus analysis (Figure 1). In the context of academic English teaching, a combination of formal and functional analysis of academic formula can help learners acquire programmatic language[9].

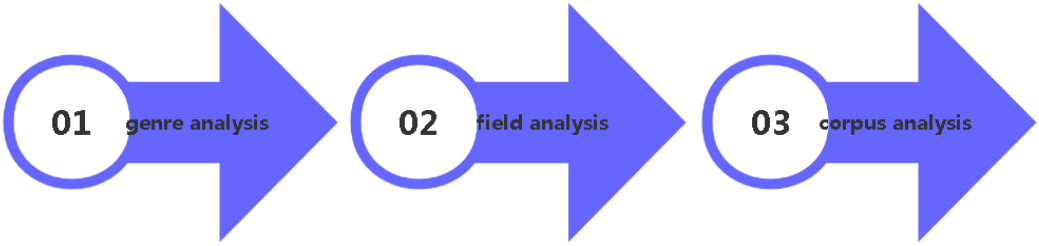


*Figure 1. "form-function" mapping of academic formula*

### 3.1. Search Methods

The corpus can be classified into heterogeneous, homogeneous, systematic and specialized types according to their contents and attributes, but these four types are too rough and not suitable for teaching. The article tends to classify corpus types from the perspective of teaching and learning in a multidimensional way so as to facilitate teachers' selection of teaching resources. For example, the corpus can be divided into spoken corpus, written corpus, native speaker corpus, learner corpus, monolingual corpus, multilingual corpus, general corpus, and specialized corpus. It should be noted that this multidimensional division is cross-cutting, and a corpus can belong to multiple types at the same time(Table 2).

*Table 2. Corpus resources for English academic formula teaching*

| Name | Website | Function |
|---|---|---|
| Linggle | https://linggle.com/ | To determine whether a multi-word pairing is accurate, you can search by word type Search for the fixed collocation of a word |
| Netspeak | https://netspeak.org/ | To determine whether a multi-word pairing is accurate, you can search by word type Search for the fixed collocation of a word |
| Just The Word | http://www.just-the-word.com/ | To determine whether a multi-word pairing is accurate, you can search by word type Search for the fixed collocation of a word |
| ANG | http://www.anc.org | 2200 Spoken and Written American English Vocabulary |
| COCA | https://www.english-corpora.org/coca/ | To retrieve the word frequency of multi-word collocations in the corpus |
| Sketch Engine for Language Learning | https://www.sketchengine.eu/skell/ | To determine whether multi-word collocations are accurate and authentic |

One of the features of Linggle is the use of "lexeme" + keyword to look up collocations. The lexical properties can be verbs v., adjectives adj., adverbs adv., nouns n., prepositions prep. and so on. One of the most useful features of Linggle is the ability to quickly find authentic multi-word expressions by using lexical qualification. For example, in Chinese we use the word "beam" to define light, such as "a beam of light", or "a ray of light", or "a line of light", but what is the English

equivalent of "light"? On linggle.com, you can use the underscore _ to replace the position of the word you are looking for. In the search box of linggle.com, type: a _ of light


*Figure 2. search result of* a _ of light

As can be seen, a beam of light, a ray of light, a flash of light, and a burst of light are all authentic usages(Figure 2).

### 3.2. Retrieval Techniques

The corpus can be searched by keyword search, fuzzy search, collocation search, sentence pattern search, pairwise word search, multiple works joint search, multiple translators joint search, and automatic ranking of search results, etc., and validation studies are conducted based on the search results. The corpus search techniques can be divided into validation search and exploratory search. Among them, the purpose of validation search is to verify whether a certain multi-word collocation is commonly used and authentic, that is, to verify whether the multi-word collocation is a program language. The purpose of exploratory search is to discover new English word collocations or sentence structures, i.e., to find out which collocations contain certain words and whether there is a program language in them. The two methods are often used together in programmatic language teaching. For validation searches, teachers can use the COCA and Sketch Engine for Language Learning (SKELL) corpora as teaching resources. Teachers first use COCA to query the frequency of a multi-word collocation in the corpus to infer whether the expression is authentic or not. For those multi-word collocations that are not "authentic", teachers can use "how to find more authentic English expressions" as an entry point to introduce the search methods and techniques of the SKELL corpus. The SKELL corpus can find all the common collocations of a word and sort them according to word frequency, which is equivalent to discovering program words and is an exploratory search. In order to verify whether the collocations found by SKELL are more authentic, COCA can be used again to verify. For exploratory search, teachers can use Linggle or netspeak to discover program words related to a word from multiple perspectives. Use * at the location where you want to insert words to search for words that are suitable for that location, or _ to search for multiple words that are suitable for that location.For example, you would like to know how to match the word “suggestion” after it. Type suggest _ in the search box, and the results will tell you the word combination of “suggestion”and the frequency of each combination(Figure 3) [12-13].


*Figure 3. search result of* suggest

Based on the word frequency, you can find out the common program words used by native speakers. Teachers can also use AntConc (https:// www.laurenceanthony.net/software/antconc/) to conduct an exploratory search of the text corpus downloaded to the local language [15].

(2) Teaching examples

Example 1: Verify which of the four phrases "few researches, few studies, little research, little study" is the most authentic expression for "few studies". First, the search box of COCA is used to check the frequency of the above four phrases. From the search results, we can see that few researches has a total of 3 corpora, few studies has a total of 1296 corpora, little research has a total of 1295 corpora, and little study has a total of 82 corpora. The three corpus of new researches are from ACAD: Romanian Economic and Business Review; ACAD: KSII Transactions on Internet; ACAD: Journal of Instructional Psychology; and 1292 corpus of new studies. Psychology; 1296 for new studies are mostly from Web, Blog, Magazines, News; 1295 for little research are mostly from Web, Blog, TV, Movie; 82 for little study are mostly from Spoken, Blog, ACAD. Finally, by comparing the word frequencies, we can conclude that few studies and little research are the more common and authentic expressions of "few studies". Also, because of the high cooccurrence of the words in these two phrases, they can be identified as program words. Also, it can be found that study as "research" is a countable noun, while research as "study" is an uncountable noun.

Example 2: Using the corpus retrieval tool AntConc to explore lexical collocations and sentence structures. The significant role of corpora for teaching and learning is to provide realistic examples and contexts, to quantify the search results, and to provide a more visual observation of language use. In the main Antconc interface, there are powerful tools such as 'Concordance Plot', 'FileView', 'Cluster', ' Collocates', 'WordList', 'Keywordlist', etc. These features can assist users in discovering certain language rules and patterns, which complement the inquiry-based learning model these features help users discover certain language rules and patterns, which complement the inquiry-based learning model. First of all, Concordance, for example, can distinguish near-synonyms, summarize the semantic rhyme or semantic tendency of words, generalize to the grammatical level of class connection, and also select the usage of collocations in different positions, such as take*of. The retrieved examples can be exported as exercises or tests to consolidate the learned points. Cluster or N-gram can summarize the common collocations or

idioms with a certain word; KeywordList and Wordlist can be combined to derive the high frequency words of a certain corpus (e.g. an article in a university English textbook) relative to other corpus, which can be used for text content and Genre analysis. In addition, Antconc can reveal the context of the text and can calculate the strength of collocations, which is more convincing with real examples of sentences.

## Funding

This work was supported by The Fund Project of Jiangxi Social Science Planning Project (19YY10).

## Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

## Conflict of Interest

The author states that this article has no conflict of interest.

## References

*[1] Ackermann K & Chen Y-H. Developing the Academic Collocation List (ACL) - A corpus-driven and expert judged approach. Journal of English for Academic Purposes, 2013, 12(4) : 235-247. https://doi.org/10.1016/j.jeap.2013.08.002*

*[2] Durrant P. Investigating the viability of a collocation list for students of English for academic purposes. English for Specific Purposes, 2009, 28(3): 157-169. https://doi.org/10.1016/j.esp.2009.02.002*

*[3] Durrant P. Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. Applied Linguistics, 2015, 38(2): 165-193. https://doi.org/10.1093/applin/amv011*

*[4] Durrant P & Mathews-Aydnl J. A function-first approach to identifying formulaic language in academic writing. English for Specific Purposes, 2011, 30(1): 58-72. https://doi.org/10.1016/j.esp.2010.05.002*

*[5] Handl, S & E. M. Graf. Collocation, anchoring, and the mental lexicon—An ontogenetic perspective. In H. J. Schmidt & S. Handl (eds.). Cognitive Foundations of Linguistic Usage Patterns. Berlin/New York: Walter de Gruyter Mouton,2010:119. https://doi.org/10.1515/9783110216035.119*

*[6] Liu Shiming, Fang Mang. A Study of Programmatic Language Types in Second Language Acquisition. Language and Culture Education Research, 2002,(2):24- 26.*

*[7] Lu Xiaofei, Liu Yingying. Research and teaching application of academic English program language based on corpus. Foreign Language, 2019(5):30.*

*[8] Moreno A I & Swales J M. Strengthening move analysis methodology towards bridging the function-form gap. English for Specific Purposes, 2018, 50: 40-63. https://doi.org/10.1016/j.esp.2017.11.006*

*[9] Nattinger, J. R. and J. S. DeCarrio. Lexical Phrases and Language Teaching[M]. Oxford: Oxford University Press, 1992.*

*[10] Perera, N. S. The Role of Prefabricated Language in Young Children's Second Language Acquisition. Unpublished Ph. D. Dissertation in Georgetown University, 2001. https://doi.org/10.1080/15235882.2001.10162797*

*[11] Yin Bangyan. Research on the structure of English sets. Journal of the PLA Foreign Language Institute, 1996, (2):1- 8.*

*[12] Weinert, R. The role of formulaic language in second language acquisition: a review. Applied Linguistics, 1995, 16: 181- 205.*

*[13] Widdowson, H. Aspects ofLanguage Teaching. Oxford: Oxford University Press, 1990.*

*[14] Wray, A. & M. R Perkins. The function of formulaic language: An integrated model. Language and Communication, 2000, 20(1). https://doi.org/10.1016/S0271-5309(99)00015-4*

*[15] Wray, A. Formulaic Language and the Lexicon. Cambridge: Cambridge University Press, 2002:332,xi. https://doi.org/10.1017/CBO9780511519772*