孙海燕◎著

英语名词搭配发展特征研究

An Analysis of Chinese EFL Learners' Development of Noun Collocation

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

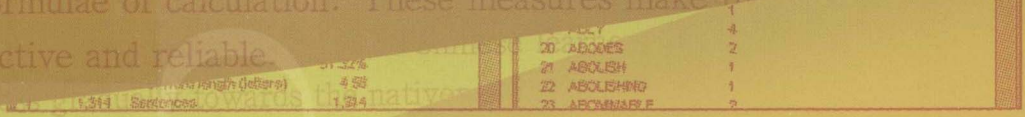

Figure 4.2 Sample screens from WordSmith of concordances and wordlists

当代语言学研究文库
河南师范大学学术专著出版基金资助项目
教育部人文社会科学研究青年基金项目

英语名词搭配发展特征研究

An Analysis of Chinese EFL Learners' Development of Noun Collocation

孙海燕 著

上海交通大学出版社

内容提要

本书基于英语学习者语料库的真实数据，采用中介语对比分析方法，考察不同阶段中国英语学习者名词搭配的发展特征。书中以名词搭配为中心，探讨英语学习者在类联接、搭配和词串方面的发展特点，分析学习者在语言使用中的句法、语义和语用特征。本书在赋码语料库的基础上分析学习者名词类联接的发展模式，从词语搭配和语法搭配方面探讨名词搭配的发展特征，从语用角度研究学习者词串的使用特点，加深对学习者语言习得路径的了解。

本书适合高等院校语言学专业的研究生、英语教师和研究者阅读参考。

图书在版编目(CIP)数据

英语名词搭配发展特征研究/孙海燕著. —上海：上海交通大学出版社，2013
(当代语言学研究文库)
ISBN 978-7-313-09604-3

Ⅰ. 英… Ⅱ. 孙… Ⅲ. 英语—名词—研究 Ⅳ. H314.2

中国版本图书馆 CIP 数据核字(2013)第074…号

英语名词搭配发展特征研究

著　　者：孙海燕
出版发行：上海交通大学出版社　　地　　址：上海市番禺路 951 号
邮政编码：200030　　电　　话：021-64071208
出 版 人：韩建民
印　　制：常熟市大宏印刷有限公司　　经　　销：全国新华书店
开　　本：787mm×960mm　1/16　　印　　张：13.5
字　　数：252 千字
版　　次：2013 年 11 月第 1 版　　印　　次：2013 年 11 月第 1 次印刷
书　　号：ISBN 978-7-313-09604-3/H
定　　价：43.00

前言

搭配是语言使用中的普遍现象，掌握搭配有助于学习者产出地道的英语。类联接作为更高一级的词类之间的搭配，可以体现词语的语法型式。语料库语言学的蓬勃发展以及学习者语料库在外语教学与研究中的应用，使我们能更客观全面地考察学生的语言特征。

本书基于学习者语料库提供的真实数据，采用中介语对比分析方法，探讨不同阶段学习者使用名词搭配的发展特征。语料库取自中国学习者英语语料库中的三个子语料库：①高中生英语作文子语料库；②大学非英语专业1～2年级学生作文子语料库；③大学英语专业3～4年级学生作文子语料库。参照语料库LOCNESS由本族语大学生的议论文组成。研究利用各种软件来提取信息，如PowerGREP，WordSmith，使用赋码软件WinBrill对语料库进行词类赋码，采用计算Z值的统计方法来确定搭配词的显著程度。本书结合了定量分析和定性分析的研究方法，主要研究内容和结果包括以下三个方面：

第一，在赋码语料库的基础上探索学习者名词类联接的发展模式。具体分析了八种名词类联接：①名词＋介词；②介词＋名词；③名词＋that从句；④名词＋不定式；⑤动词＋名词；⑥名词＋动词；⑦形容词＋名词；⑧名词＋名词。研究发现，学习者过少使用“名词＋介词”和“名词＋that从句”；而过多使用“名词＋名词”。本书参照MacWhinney的“竞争模式”，根据汉英语言结构的差异分析了母语迁移在形成上述模式中的作用。由于汉语结构的影响，学生倾向于过少使用与汉语不同的结构，而过多使用和汉语相近的结构。

第二，从词语搭配和语法搭配方面探讨学习者名词搭配的发展特点。研究发现，三组学生的搭配水平逐渐发展，搭配的语义精确性和句法复杂性随着学生语言水平的提高而改进。具体来说，在词语搭配方面，通过分析节点词“problem，situation，future”的搭配方式考察学习者的词语搭配能力。数据表明三组学生的搭配水平有不同程度的进步；但

是和本族语者相比，中国学生使用大量不地道的词语搭配，并倾向于使用较早习得的高频词作为搭配词。在语法搭配方面，以“idea，reason，fact”为节点词，通过提取索引行分析学习者语法能力的发展特征。语法搭配反映语言使用的复杂性，研究发现高水平学习者使用较复杂的语言结构。尽管学习者的搭配知识总体上呈发展趋势，石化现象依然存在，在词语搭配方面尤为显著。本书分析了学生在词语搭配上产生问题的原因，即母语迁移、迂回策略、词汇知识不全面和搭配意识的缺乏。关于如何提高学生的搭配能力，本书提出了教学中搭配选择的标准，建议在语境中学习词汇，鼓励自主学习并培养学生的搭配意识。

第三，从语用角度分析学习者使用的词串，考察三组学生使用词串的发展特征。词串作为搭配的一种延伸，是学习者语言知识不可或缺的组成要素。通过分析语料库中节点词“way，part，place”所组成的词串，研究发现，高中组学生主要依赖词语的指示意义，缺乏语用能力；大学非英语专业1～2年级学生能较好地掌握词语的功能意义，从而有较高的语用能力；英语专业3～4年级学生语用能力最高。换言之，较高水平学习者善于使用词串实现组织语篇和表达态度等各种功能，其产出的词串更接近本族语者使用的典型词串。但是和本族语者相比，中国学习者在词串使用的频数、种类和功能方面都不尽相同。中国学生不擅长使用具有语用功能的词串，其语用能力的缺乏可能是由于在教学过程中，词汇的语义得到强调而功能往往被忽视所造成的。

本书具有一定的理论意义和方法论意义，并对教学实践具有较好的指导作用。本书从句法复杂性、语义精确性和语用功能方面综合研究学习者语言的发展特征，使我们深入全面地了解中介语的发展过程。在研究方法上，语料库数据为研究学习者语言提供了有力的数据支持，综合利用赋码软件、检索软件和统计方法考察中介语，为今后的研究提供了新的思路。在教学实践中，大纲设计人员、教材编写人员、教师与学生应该认识到学习者搭配能力的发展特点，英语教学应根据不同阶段学生的发展特征和需求进行安排。

本书是教育部人文社会科学研究青年基金项目（编号：12YJC740088）和河南省高校科技创新人才支持计划（人文社科类）（编号：[2012]469号）的研究成果。

在本书付梓出版之际，我谨向所有关心、帮助过我的人们表示真挚的谢意。首先要感谢我的导师陈永捷教授。他严谨的治学态度、广博的知识、精益求精的工作作风使我受益匪浅，正是在他的悉心指导和亲切关怀下，我的学业得以顺利完成。

在我攻读博士以及修改书稿期间，许多学者给我提供了无私帮助。感谢上海交通大学的所有授业恩师，卫乃兴教授指引我踏入语料库语言学的研究领域，俞理明教授、周国强教授、王同顺教授、潘文国教授对研究提出了宝贵的建议。感谢北京外国语大学的李文中教授和梁茂成教授，他们关于语料库软件操作的指导开阔了我的研究思路。

上海交通大学的学友和河南师范大学的同事给了我热情的帮助。感谢李素枝博士、崔艳嫣博士、赵勇博士、常辉博士、甄凤超博士等好友的鼎力相助。

最后，我要感谢全家对我的支持。特别感谢我的丈夫齐建晓先生，他无微不至的关怀、一如既往的鼓励是我学业进步的源泉。

由于作者水平有限，书中疏漏和错误在所难免。不足之处，恳请各位专家同行批评指正。

孙海燕

河南师范大学外国语学院

2013 年 3 月

Contents

Chapter 1 Introduction

1.1 Research Background

Corpus linguistics offers a large quantity of authentic data for linguistic studies and many issues are amenable to corpus-based research. Corpus-based analyses provide profound insights into various areas of language structure and use by utilizing corpora in conjunction with computational techniques. With the great amount of evidence derived from corpora, the unity of meaning and pattern has recently been recognized (Sinclair 1991). Text studies and corpus studies have revealed the intricacy of the links between words, for example, their strong clustering tendencies and the patterns which are associated with them. Hunston and Francis (2000) argue that it is of no avail to deal with syntax and lexis separately because they are interdependent. A strong association is believed to exist between lexis and grammar. Based on this belief, the present study attempts to investigate two aspects of language use—colligation and collocation, which integrate grammar and lexis to some extent.

Skehan (1998) proposes that representation functions by means of a dual-mode system, with access to rules and exemplars. The rule-based system is likely to be parsimoniously and elegantly organized, with rules being compactly structured. It prioritizes analyzability, but its operation will lead to a heavy processing burden during ongoing language use. The exemplar-based system, however, is primarily based on the operation of a redundant memory system in which there are multiple representations of the same lexical elements, and as a result, the system lacks parsimony. But the gain of such a system is processing speed in that utterance units do not require excessive internal computation. It is argued that neither the rule-based system nor the exemplar system is ideal independently. The former leads to the development of an open, form-oriented system, while the latter emphasizes meaning and is less appropriate for underlying system change. In this research, the use of colligation is regarded as a reflection of rule-based system while collocation is treated as one type of exemplars. On the basis of this dual-mode system, this

book aims to investigate how Chinese learners represent the rule-based system and the exemplar-based system through the examination of their use of colligations and collocations respectively.

Colligation refers to a generalizable class of collocation, and its construct is specified by word classes rather than as distinct lexical items. Examining second language (L2) learners' use of colligations offers a description of their language production in terms of structurally defined patterns. There has been scanty exploration, however, into L2 learners' use of colligations. And the prior studies were almost exclusively conducted in light of the definition proposed by Lewis (2000), focusing on the interaction between the particular lexical items and the grammatical patterns they formed. To the knowledge of the present author, the study of L2 learners' use of colligations on the basis of tagged corpus remains blank, probably because part of speech (POS) taggers are not readily available or not easy to operate. To bridge the gap, the present research on colligation operates purely on the level of word class, which is entirely different from the studies carried out previously.

Collocation is the way words combine in a language to produce natural-sounding speech and writing (Crowther et al. 2001). Collocation has often been considered a problematic area for English as a foreign language (EFL) learners (Bahns & Eldaw 1993; Channell 1981; Granger 1998b; Howarth 1998). In the process of vocabulary learning, L2 learners always deem it their main task to build up a large repertoire of vocabulary, without giving due attention to the typical collocational behavior of words. In consequence, they frequently make grammatically well-formed sentences which nevertheless sound awkward or unnatural to the native ears. A brief glance at the relevant literature indicates that there has been a mushrooming amount of research on collocation since 1980s. Miscellaneous findings have been obtained from the analysis of L2 learners' collocational behavior, from which much insight has been gained concerning how learners acquire collocations and how to teach collocations more effectively. Nevertheless, little has been conducted to investigate L2 learners' collocational knowledge from a developmental point of view. In addition, most previous collocational studies focus on the verbal behavior (e.g. Chi et al. 1994; Howarth 1998; Nesselhauf 2003; Pu 2000), while the research in the area of nouns is very much in its infancy. The pivotal role of nouns has come to be recognized by researchers with the assumption that the

pattern of collocations is usually triggered off by nouns and nouns are the most suitable headwords for collocation searches (Crowther et al. 2001; Woolard 2000). The investigation into the area of nouns remains quite sparse in spite of the essential part played by them in information building. In view of the great importance of nouns and the paucity of developmental studies, this book intends to examine Chinese EFL learners' behavior of noun collocations across three proficiency levels, with the purpose of revealing the developmental features of Chinese learners' collocational competence.

To extend the study of collocations, the characteristics of Chinese learners' use of clusters are explored. Clusters are continuous strings of words occurring repeatedly in identical form. As essential building blocks to convey language users' intentions and reactions in discourse, they can be exploited to achieve socio-interactional functions (Alternberg 1998; Wray 2000). This research investigates the 3-word clusters produced by Chinese learners, analyzing the meanings of the clusters and their pragmatic functions. It is hoped that the investigation can shed light on the development of pragmatic competence of Chinese learners.

In a nutshell, the present research is oriented towards an exploration into the developmental features of Chinese EFL learners' use of noun colligations, collocations and clusters. The factors exerting influence on them are also dealt with. It is anticipated that by integrating these three aspects of language performance in the present study, a full account of Chinese learners' use of nouns can be achieved.

1.2 Research Objectives and Questions

The central purpose of this research is to investigate the characteristics of noun colligations, collocations and clusters produced by Chinese EFL learners. A corpus-based, cross-sectional study is carried out to examine Chinese learners' performance across three proficiency levels: (1) senior high school learners, (2) college non-English majors (Year 1-2), (3) English majors at tertiary level (Year 3-4). It is assumed that they represent three consecutive stages of interlanguage development: beginning, intermediate, and advanced.

This book aims to explore the developmental features of Chinese learners' use of noun colligations, collocations and clusters. Specifically, the analysis of colligation is conducted through the investigation of the co-occurrence of word class on the

basis of POS tagged corpora. Then the learners' collocational behavior is examined, focusing on the inappropriate collocations with reference to the native speaker (NS) corpus to testify the idiosyncratic features of the collocations employed by Chinese learners. The taxonomy by Benson, Benson and Ilson (1986) is followed, that is, English collocations are classified into two major groups: lexical and grammatical collocations. To capture the totality of the learners' knowledge of word combinations, the recurrent strings of clusters used by the learners are retrieved from the corpora and examined with regard to their meanings as well as functions. Therefore, the objectives of the present research are threefold. First, this study attempts to probe into the developmental pattern of Chinese EFL learners' use of noun colligations. Second, this research aims to explore the developmental features of Chinese learners' behavior of lexical and grammatical collocations. Third, the present study intends to examine the meaning aspects and the pragmatic functions of the clusters produced by three groups of Chinese learners. In addition, the reasons that account for these features are touched upon.

In light of the research objectives stated above, three major questions are addressed:

(1) What is the developmental pattern of Chinese learners' use of noun colligations across three proficiency levels?

(2) What are the developmental features of Chinese learners' behavior of lexical collocations and grammatical collocations?

(3) What is the use of clusters characterized by the three groups of Chinese learners?

1.3 Organization of the Book

The book is composed of 8 chapters. Chapter 1 serves as a brief introduction to the research background, research objectives and questions. Chapter 2 reviews the literature on corpus linguistics, prefabricated language, and most importantly, the theoretical exploration and empirical investigation into colligation, collocation and cluster. Chapter 3 deals with the different approaches to the study of learner language—contrastive analysis (CA), error analysis (EA) and contrastive interlanguage analysis (CIA). Chapter 4 discusses the methodology of a corpus-based, cross-sectional empirical study and the various tools exploited, including the statistical

tool of SPSS, the tagging tool of WinBrill, and the concordance software of PowerGREP, TACT, MicroConcord and WordSmith. The findings pertaining to the research questions are reported and discussed in the following three chapters. Chapter 5 first presents a framework of the eight types of colligations under study, then probes into the developmental pattern of the noun colligations employed by Chinese learners across three proficiency levels on the basis of POS tagged learner corpora. Chapter 6 traces the development of lexical collocations and grammatical collocations by a wealth of detailed analyses, with an emphasis on the varying degrees of semantic accuracy and syntactic complexity of the collocations produced by the three groups of Chinese learners. It identifies the sources of collocational problems, and puts forward corresponding proposals to enhance the learners' collocational competence. Chapter 7 analyzes the 3-word clusters retrieved from the corpora in terms of frequency, variety, and function in an attempt to identify the distinctive features of clusters used by Chinese learners. Chapter 8 brings together the major findings of the research, and expands upon the theoretical, methodological and pedagogical implications. Limitations of the present research and directions for future research are also discussed.

Chapter 2 Corpus Linguistics, Collocation and Colligation

2.1 Introduction

The studies reviewed in this chapter fall into four major realms. First, a general overview of corpus linguistics is presented, and in particular, the benefits of computer learner corpus are addressed. Second, the notion of prefabricated language and its importance to language learning is introduced. Prefabricated language has become a major focus of interest in English language teaching (ELT) because there is a general recognition of the problem facing L2 learners in achieving the naturalness of native-speaker use that derives from the appropriate selection of conventional phraseology (Hakuta 1974; Howarth 1998; Nattinger & DeCarrico 1992). Third, the theoretical exploration and empirical investigation into collocations as well as colligations are reviewed. The fourth section deals with the related studies of word clusters. As an extension of collocations, clusters are of paramount significance in acquisition and communication because of their high frequency in language use.

2.2 Corpus Linguistics

2.2.1 Overview of Corpus Linguistics

Corpus linguistics is "the study of language based on examples of 'real life' language use" (McEnery & Wilson 1996: 1). To fully understand what corpus linguistics is entails defining the term "corpus". According to Sinclair (1991: 171), a corpus refers to "a collection of naturally occurring language text, chosen to characterize a state or variety of a language". Francis (1992: 17) defines it as "a collection of texts assumed to be representative of a given language, dialect, or other subset of a language to be used for linguistic analysis". From these definitions

it can be seen that corpus data should be naturally-occurring and the selected texts need to be representative of a given language. More recently, the term has been reserved for collections of texts that are stored and accessed electronically. Because computers can hold and process large amounts of information, electronic corpora are usually larger than the paper-based collections previously used to study language. Taking this aspect into account, Tognini-Bonelli (2001: 55) defines corpus as follows.

> A corpus is taken to be a computerized collection of authentic texts, amenable to automatic or semiautomatic processing or analysis. The texts are selected according to explicit criteria in order to capture the regularities of a language, a language variety or a sub-language.

It is the computer that allows us to exploit corpora on a large scale with speed and accuracy. As Kennedy (1998: 5) points out, "corpus linguistics is thus now inextricably linked to the computer, which has introduced incredible speed, total accountability, accurate replicability, statistical reliability and the ability to handle huge amounts of data". Several advantages of the corpus-based study come from the use of computers and the automatic processing techniques. First, computers make it possible to identify and analyze complex patterns of language use, allowing the storage and analysis of a larger database of natural language than the former one which could be dealt with by hand. In fact, analysis of large corpora or of many complex features simply would not be feasible without a computer. Second, computers enable us to perform such linguistic research as multi-dimensional analysis, cluster analysis, factor analysis, etc., with the exploitation of multivariate techniques. Third, the machine-readable corpora can be easily enriched with extra information. Statistical and probabilistic information can be obtained through concordancing packages and programs, which are very useful for investigating word frequency, type-token ratio, average word length, average sentence length, and so on.

Owing to the widespread use of computerized corpora among teachers and researchers alike, there is a growing expectation that description of language will be based on quantities of authentic data rather than on a linguist's intuitions and/or prejudices (Hunston & Francis 1998). Corpus linguistics focuses on a more

empiricist, rather than rationalist view of scientific enquiry. Empirical data enable the linguists to make objective statements, rather than those that are subjective, or based upon the individual's own perception of language. Biber, Conrad and Reppen (1998: 4) summarize the essential characteristics of corpus-based analysis as follows:

> (1) it is empirical, analyzing the actual patterns of use in natural texts;
> (2) it utilizes a large and principled collection of natural texts, known as a "corpus", as the basis for analysis;
> (3) it makes extensive use of computers for analysis, using both automatic and interactive techniques;
> (4) it depends on both quantitative and qualitative analytical techniques.
>
> "Taken together, these characteristics are advantageous for linguistic studies, resulting in a scope and reliability of analysis not otherwise possible" (ibid.).

Despite the advantages derived from the exploitation of corpus, the role of corpus linguistics in language study has given rise to much controversy. McEnery and Wilson (1996) maintain that corpus linguistics is a methodology rather than an aspect that needs to be described; it serves as a means of verifying hypotheses about a language. They explain that it might not be considered as a branch of linguistics in the same sense as syntax, semantics, sociolinguistics and so forth. In contrast, Tognini-Bonelli (2001: 48) asserts that corpus linguistics is much more than just a methodology since it leads to the identification of a new unit of currency for linguistic description and radically affects the way in which languages are described and theories about them constructed:

> What had started as a methodological enhancement but included a quantitative explosion has turned out to be a theoretical and qualitative revolution in that it has offered insights into the language, that have shaken the underlying assumptions behind many well established theoretical positions in the field.

Tognini-Bonelli makes a distinction between corpus-based approach and corpus-driven approach. The corpus-based approach starts with a set of explicit rules and validates these statements by using corpus data. It refers to a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions. Corpus evidence is not regarded as a determining factor with respect to the analysis, which is carried out according to pre-existing categories. The corpus-driven approach, by contrast, builds up the theory step by step in the presence of evidence. The observation of certain patterns leads to a hypothesis, which in turn leads to the generalization in terms of rules of usage and finally finds unification in a theoretical statement. It can be seen as primarily inductive because it does not start with an openly stated rule but derives it by generalizing from particular language facts. Such a distinction illuminates the important differences between these two types of study. Though corpus-based research is prevalent currently, corpus-driven approach may begin to prosper in future linguistic study.

2.2.2 Computer Learner Corpus

One type of corpus that specifically focuses on the teaching process and is particularly useful for error analysis is what is referred to as a learner corpus (Tognini-Bonelli 2001: 9). Computer learner corpora (CLC) are "electronic collections of authentic texts produced by foreign or second language learners" (Granger 2003: 538) on the basis of certain explicit criteria and for a specific purpose.

Second language acquisition (SLA)[①] research employs a variety of data types. Ellis (1994: 670) classifies them into the following categories: (1) language data, which reflect learners' attempt to use the L2 in either comprehension or production; (2) metalingual judgments, which tap learners' intuitions about the L2; and (3) self-report data, which explore learners' strategies via questionnaires or think-aloud tasks. Language data are said to be "natural" if no control is exerted on the learners' performance and "elicited" if they result from a controlled

① In this study, no strict distinction is made between second language acquisition and foreign language learning. Just as Ellis (1985: 5) maintains, "second language acquisition is not intended to contrast with foreign language acquisition".