大数据科学研究丛书

Structural Equation Modeling in Applied Linguistics in the Era of Big Data

大数据时代应用语言学研究中的结构方程建模

王天剑　王彦之　编著

科学出版社

大数据科学研究丛书

Structural Equation Modeling in Applied Linguistics in the Era of Big Data

大数据时代应用语言学研究中的结构方程建模

王天剑　王彦之　编著

科学出版社

北京

内 容 简 介

传统应用语言学研究主要涉及结构化数据（即小数据）。随着信息技术发展，应用语言学研究也进入了大数据时代。这里的大数据是指应用语言学研究使用的超级语料库，它们具有一般大数据的基本特征。本书属于统计学和语料库语言学交叉学科范畴，着重考察三个方面：①语料库大数据的特征；②常用的语料库分析软件及其在语言数据提取中的应用；③结构方程建模的概念和原理及其利用语料库数据建模的基本方法。

作为一部交叉学科著作，本书可供具有一般英语阅读能力，且对大数据语料库语言学和结构方程建模感兴趣的教师、大学生或科研工作者作为参考性工具书或者教材使用。

图书在版编目（CIP）数据

大数据时代应用语言学研究中的结构方程建模 = Structural Equation Modeling in Applied Linguistics in the Era of Big Data：英文 / 王天剑，王彦之编著. —北京：科学出版社，2019.11

（大数据科学研究丛书）

ISBN 978-7-03-062949-4

Ⅰ. ①大… Ⅱ. ①王… ②王… Ⅲ. ①应用语言学－数据处理－系统建模－研究－英文 Ⅳ. ①H08 ②TP391.92

中国版本图书馆 CIP 数据核字（2019）第 241015 号

责任编辑：马 跃 李 嘉 / 责任校对：杨 赛
责任印制：张 伟 / 封面设计：无极书装

科 学 出 版 社 出版
北京东黄城根北街 16 号
邮政编码：100717
http：//www.sciencep.com

北京建宏印刷有限公司印刷
科学出版社发行 各地新华书店经销
*
2019 年 11 月第 一 版 开本：720 × 1000 B5
2019 年 11 月第一次印刷 印张：11 1/4
字数：226 000

定价：92.00 元

（如有印装质量问题，我社负责调换）

作者简介

王天剑（1968—），汉族，籍贯河南南阳，贵州财经大学外语学院教授，海归博士。近年来在《外语学刊》《外语与外语教学》《中国特殊教育》《中小学英语教学与研究》《中国应用语言学》等刊物上发表论文20余篇，其中2篇被中国人民大学复印资料全文转载；专著《语言焦虑研究入门》（英文，中国社会科学出版社出版）荣获贵州省第十次哲学社会科学奖三等奖。

王彦之（1993—），汉族，籍贯河南南阳，中南大学地球科学与信息物理学院取得硕士学位。近年来在《鲁东大学学报（自然科学版）》《上海工程技术大学学报》《四川理工学院学报（自然科学版）》《水文地质工程地质》等刊物上发表学术论文多篇。

Preface

"Big data" is a label for data sets which are too complex or large to be processed by traditional applications. Challenges range from search and capture to sharing, storage, transfer and analysis. "Big data" is generally related to the employment of predictive analytics or other approaches to extract valuable information from large data sets. It is producing a growing effect on just everything including government, business, science, research, and even our mind.

The book provides clear guidance on how to correctly conduct research using Structural Equation Modeling (SEM) in applied linguistics in the era of big data. The primary purpose is to help new researchers to overcome fears and to encourage them to start the practice of SEM with such data sets.

The book presents SEM in a clear and accessible way for common learners. It focuses on practical methods rather than statistical theories. Even without broad statistical knowledge, readers and applied researchers will find it easy to understand and use SEM with big data. If you are enthusiastic about the technique, you will find the book enlightening.

The book covers various key techniques in SEM with big data, such as the transformation of big data into traditional data, path analysis, factor analysis, structural regression analysis, analysis with order-categorical data, analysis with censored data, data imputation, model comparison and estimation method selection.

The book is organized in line with how readers generally progress with the learning of statistical analysis. General introduction and fundamental concepts are followed by specific techniques; basic techniques are followed by more complex ones.

Simplified examples are provided to illustrate the application of SEM with big data. These examples are mostly adapted from the author's personal research in applied linguistics. Problematic examples are created and included to display the ways to deal with various problems that you may meet in real studies. Plenty of figures and tables are employed to illustrate the steps and procedures of analyses, as well as the output in text or graphic formats. Readers can copy those example data to their own computers and practice the central techniques on their own computers.

The book supports formal course teaching, as well as the self-study of SEM with big data. In any case, you will enjoy the process of learning by doing.

With the inclusion of big data, the book is an update to the similar previous books on SEM. Your knowledge about the vast possibilities of the technique will be greatly expanded. The book can also be applied to the researches in other social and behavioral sciences besides linguistics.

The book is dedicated to the use of Amos, though several other programs are also presently popularly used in SEM, such as EQS, LISREL and Mplus. The use of those programs can be easily grasped once you are familiar with Amos.

Contents

Chapter 1 Introduction

Applied linguistics is a branch of linguistics, which focuses on the teaching and learning of language. Quantitative research in applied linguistics has to take advantage of many statistical techniques, one of which is SEM. In the era of big data, the use of SEM in applied linguistics becomes more complicated. This chapter provides an introduction to the fundamental concepts of the book, the use of big data corpora resources, and SEM with the software Amos. The introduction is supposed to help you understand what is to come in later chapters.

1.1 Applied Linguistics and SEM in Applied Linguistics

1.1.1 Applied Linguistics

Established in the mid 20th century, applied linguistics is an interdisciplinary subject. It is a science concerned with the application of linguistic theories to various fields of studies. It identifies language-relevant problems in real life and attempts to provide solutions to those problems. It is closely related to sociology, anthropology, psychology, communication research, computer science, and education. Pragmatics, discourse analysis, conversation analysis, computer-mediated communication, second language acquisition, sign linguistics and stylistics are all specific branches of applied linguistics.

Applied linguistics first came into being to distinguish itself from general linguistics in the 1950s (Davies & Elder, 2004: 1). Though initiated by experts in Europe, applied linguistics quickly became popular and flourished all over the world. It has always been concerned with language problems, among which is the problem of language teaching and acquisition. In America, Leonard Bloomfield was an early advocator of applying linguistic theories to language teaching. He laid the foundation for the Army Specialized Training Program during World War Ⅱ. Another advocator was Charles Fries, the founder of the English Language Institute at the University of Michigan. In England, the British Association for Applied Linguistics came into being

in 1967, and one of its central missions is promoting the study of language teaching and acquisition (Davies & Elder, 2004: 7). Nowadays, language teaching and learning problems have become a global focus of applied linguistics. Narrowly speaking, applied linguistics has been accepted as a professional term for the study of language teaching and acquisition. This book simply follows the trend and adopts such an interpretation of applied linguistics.

Applied linguistics involves many teaching, learning and environmental variables. These variables constitute a complicated hierarchy, affecting the process of language acquisition. Statistical techniques have been widely employed in applied linguistics. To language teachers in China, SEM is a relatively new and challenging statistical method.

1.1.2 SEM in Applied Linguistics

1.1.2.1 Concept of SEM

The term SEM is not coined to refer to a specific technique, but to designate a family of statistical techniques, including correlation, regression, factor analysis, path analysis, structural regression, and so on.

SEM is also called covariance structure analysis, analysis of covariance structures, or causal modeling. But the term "causal" can be misleading. SEM is essentially based on correlation or covariance. When A and B are correlated, you may successfully develop a model in which A is assumed to cause B, but you may as well develop a model in which B is assumed to cause A. So the model cannot decide causal relationships. It can, at most, provide statistical support for a theoretically sound hypothesis.

The origin of SEM dates back to 1904, when Charles Spearman developed the exploratory factor analysis. Later in 1918, Sewall Wright developed path analysis. In 1970s, Karl Gustav Jöreskog et al. combined path analysis and factor analysis into the same statistical system (now called SEM) and developed the computer software LISREL to do the analysis (Kline, 2011: 15). From then on, more and more researchers have improved SEM and developed better software, which has stimulated the widespread use of the technique.

SEM offers researchers an array of methods for examining relationships, testing hypotheses and developing theories. Latent variables can be included in the model to represent underlying factors of observed variables, and measurement errors or influences from unknown variables can be estimated in the analysis (Raykov & Marcoulides, 2006: 7).

1.1.2.2 Uses of SEM in applied linguistics

SEM has long been used in various branches of science. Its use in applied linguistics is relatively new. The following are examples of various relationships studied with SEM. The arrow "→" indicates the direction of prediction or influence.

Pulido and Hambrick (2008: 164) used a sample of 99 adult English-speaking learners of Spanish to examine the relationships among language processing experience, passage sight vocabulary, text comprehension, vocabulary retention, etc. The results from SEM suggested that language processing experience affected passage sight vocabulary which further influenced text comprehension. Text comprehension had an effect on vocabulary retention and growth. The modeling can be summarized as: language experience → sight vocabulary → text comprehension → vocabulary retention and growth.

Guo and Roehrig (2011: 42) investigated the English vocabulary, English syntactic awareness, general reading strategies (meta-cognitive awareness), and English reading comprehension with a sample of 278 adult Chinese EFL learners. The authors used English vocabulary and English syntactic awareness as indicators of the factor L2 knowledge, and reading strategies as the indicator of general reading knowledge. SEM shows that the two factors account for 87% of the variance of reading comprehension, with L2 knowledge being the better predictor of reading comprehension. Multiple groups analysis suggests that the results hold across both the upper and lower groups of readers. The modeling can be summarized as: L2 knowledge + general reading knowledge → reading comprehension.

To examine the respective roles of grammatical and vocabulary knowledge in reading comprehension, Zhang (2012: 558) used a sample of 190 advanced Chinese EFL learners. The result of SEM indicates that reading comprehension is significantly related to vocabulary knowledge, and weakly related to grammatical knowledge. Moreover, implicit grammatical knowledge appears to be more closely related to reading comprehension. The modeling can be summarized as: grammatical knowledge + vocabulary knowledge → reading comprehension.

In a longitudinal study, Netten, Droop and Verhoeven (2011: 413) used 729 L1 and 93 L2 learners of Dutch as participants. Reading comprehension was measured in both grade 4 and grade 6, while language proficiency, decoding, nonverbal reasoning, mathematics, reading motivation, academic self-confidence and home reading resources were all measured in grade 4. SEM is used to analyze the data for L1 and L2 participants

separately. The results reveal that, for L1 learners, 60% of the variance of reading comprehension in grade 6 could be directly or indirectly predicted by all the other variables (including the reading comprehension in grade 4). Similar results are obtained for L2 learners. The modeling can be summarized as: proficiency + decoding + reasoning + mathematics + motivation + confidence + resources → reading comprehension.

To test the Simple View of Reading, a theory which suggests that reading comprehension is the product of listening comprehension and word decoding, Verhoeven and Leeuwe (2012: 1805) drew a sample of 1,293 L1 learners and a sample of 394 L2 learners of Dutch. In grades 1, 3 and 5, listening comprehension and word decoding were tested, and in grades 2, 4 and 6, reading comprehension was tested. SEM indicated that reading comprehension could be predicted by listening comprehension and word decoding. In higher grades, the predicting effect of word decoding decreased while that of listening comprehension increased. The relationship and tendency were similar for both L1 and L2 learners, though the relationship between listening comprehension and reading comprehension seemed to be consistently stronger for L1 learners. The modeling can be summarized as: listening + decoding → reading comprehension.

Using SEM, Baker, Stoolmiller, Good Ⅲ et al. (2011: 331) analyzed the relationships among passage fluency, word reading and reading comprehension. Ninety-six second-graders participated in the study. Those children were simultaneously learning both Spanish and English, with Spanish being the home language (language used at home). The results indicate that comprehension has an influence on passage fluency for both the Spanish and the English languages even when the effect of word fluency is controlled. Though English and Spanish are two languages which differ in orthographic transparency, the relationships appear to be constant across the languages. The modeling can be summarized as: comprehension + word fluency → passage fluency.

Two hundred and twenty-seven non-English-major EFL university students were involved in the study by Wu, Yen and Marek (2011: 118), who investigated the learners' motivation, ability and confidence with questionnaires. Analysis with SEM suggests that confidence can be predicted by motivation and ability. The modeling can be summarized as: motivation + ability → confidence.

1.2 Big Data Corpora and the Extraction of Small Data

In Section 1.1, we focus on the concept of applied linguistics, and the use of SEM

in applied linguistics. But the data involved in those SEM studies can all be labeled as small data. Small data are those data which can be directly analyzed by traditional statistical techniques. In this section, we will discuss big data corpora and the extraction of small data from those corpora. The initial letters (or the initial letters of content words) of terms for buttons or options are generally in upper cases. This is applicable to cases throughout this book.

1.2.1 Concept of Big Data

"Big data" is a label for data sets which are too complex or large to be processed by traditional applications. The size of the data sets is so large that commonly used software tools are incapable to capture, manage and process within a tolerable period of time. It requires the integration of a set of techniques and technologies to extract values from datasets which are large and complex.

In a report in 2001, Gartner Group's (then Meta Group) analyst Doug Laney put forward the 3-dimensional data growth features: increasing variety, volume, and velocity (Laney, 2001). In 2012, Laney further defined big data as high-volume, high-variety and/or high-velocity information assets which need new forms of processing to promote discovery, decision making and process optimization (Laney, 2012). Currently, it is universally agreed that by means of specific technology and approach, we can extract valuable information from big data. Moreover, some experts have added new dimensions (such as "veracity" "variability") to better define the concept of big data. But anyway, it is the size which determines whether the data sets can be labeled as big data or not. Big data may include structured, unstructured or semi-structured data. To enhance the generating and processing of big data, Google Company put forward a program model called MapReduce in 2004. It has a parallel algorithm working on a cluster of loosely or tightly connected computer clusters. The program model is made up of two parts: the Map part and the Reduce part. The Map part does the filtering and classifying task, while the Reduce part does the summarizing task. The former sorts items into groups, and the latter gives a summary to the items in each group (counting the number or frequency). The MapReduce system coordinates the computer clusters in the processing of the data. The program model is so successful that its algorithm is replicated by others. In 2011, Apache Software Foundation developed an open-source software called Apache Hadoop which incorporates the functions of MapReduce. Apache Hadoop is composed of two parts: one for storing data (Hadoop Distributed File System), and

the other for processing data (MapReduce). It supports the parallel processing of big data across computer clusters, and is widely used by organizations and companies for production and study.

1.2.2 Big Data Corpora

1.2.2.1 Features of corpora

In applied linguistics research, the frequently used big data sets are corpora. A corpus is a set of text, audio, or video files (called text corpus, audio corpus and video corpus respectively). If two or three of the media are involved in a data set, it is called a multimedia corpus. Presently, techniques for coping with text corpora are well developed. In this book, the term "corpus" is to be accepted as text corpora, unless otherwise specified.

A corpus is qualified to be called big data, because it has most of the defining features of big data.

Firstly, it has a high volume. In terms of size, a corpus is large enough to be called big data. For example, the American National Corpus (ANC) has 40-50 million words. The British National Corpus (BNC) contains 100 million words.

Secondly, a corpus has a high variety of big data. A balanced corpus covers texts of all genres, and the variety is high enough to be called big data. For example, ANC includes data from court transcript, email, debate transcript, essay, government document, fiction, non-fiction, journal, newspaper, letter, spoken language, technology, travel guide, twitter, joke, blog, script, ficlet, movie spam, and other texts. BNC has similar coverage.

Thirdly, a corpus has high veracity of big data. The texts covered by a corpus are all from real life.

1.2.2.2 Wide acceptance of corpora as big data

In fact, many organizations and experts have accepted corpora as big data sets. For example, from the database of Springer Link, ScienceDirect, ProQuest, etc., if we search for "'big data' AND corpus", many records can be found (Figure 1.1-Figure 1.3).

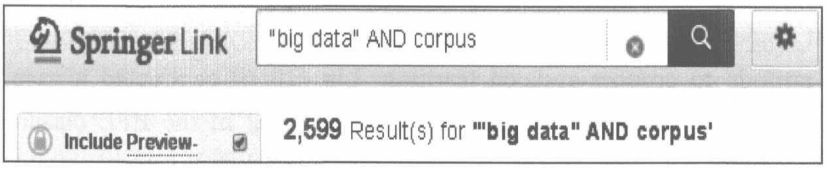

Figure 1.1 Results from searching for "'big data' AND corpus" in Springer Link

Source: Retrieved from the library database of Central South University, China [2019-05-08]

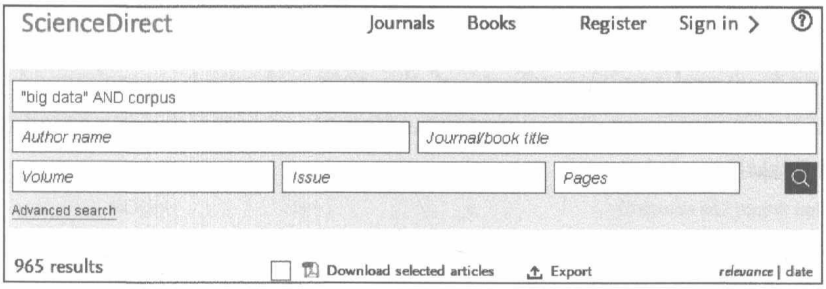

Figure 1.2 Results from searching for "'big data' AND corpus" in ScienceDirect

Source: Retrieved from the library database of Central South University, China [2019-05-08]

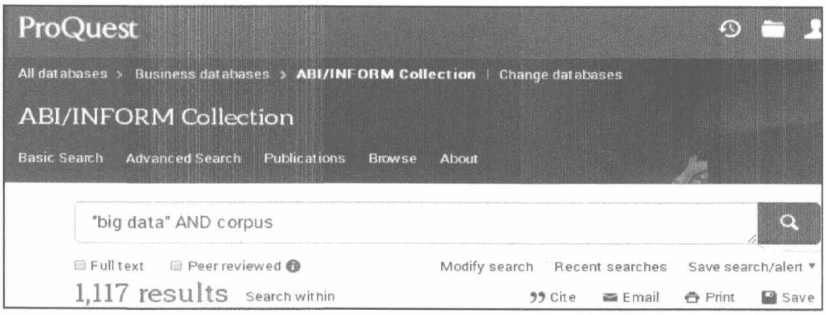

Figure 1.3 Results from searching for "'big data' AND corpus" in ProQuest

Source: Retrieved from the library database of Central South University, China [2019-05-08]

In a word, corpora have been gradually accepted as big data. In this book, we may use "corpora" or "big data corpora" alternatively. In recent years, more and more corpora have been created for linguistic researches. Some of the corpora are available online, and others can be used offline. Corresponding sorts of software have been developed to help researchers to extract valuable information from corpora.

Nowadays, many experts are using corpora in applied linguistic researches. Phoocharoensil (2012: 507) investigated Thai EFL students' attitudes to learning grammar through concordance lines (extracted from corpora). The author discovered

that most students can benefit from learning grammar by using corpora, and have a positive attitude to corpora-assisted learning. Lin (2016) conducted a case study to examine the effects of corpus-aided grammar learning in the EFL classroom. The results revealed that data-driven learning (corpus-aided learning) could enhance college students' learning attitudes in general. Furthermore, data-driven learning is considered to be innovative and interesting by the early-career teachers involved in the study. Liu and Jiang (2009: 61) explored the effects of integrating corpus and contextualized lexicogrammar in both foreign and second language learning situations. They found several positive effects of the approach, such as enhanced grasp of lexicogrammar, improved critical understanding of grammar, and increased discovery of learning skills.

1.2.3 Extracting Small Data from Online Big Data Corpora

There are presently many online corpora. Here is a list of some of the most famous ones (Table 1.1).

Table 1.1 Famous online corpora

Corpus names	Words	Time periods
Global Web-based English (GloWbE)	1.9 billion	2012-2013
Hansard Corpus	1.6 billion	1803-2005
Early English Books Online	755 million	1470s-1690s
Corpus of Contemporary American English (COCA)	560 million	1990-2017
Corpus of Historical American English (COHA)	400 million	1810-2009
Corpus of US Supreme Court Opinions	130 million	1790s-present (2017)
TIME Magazine Corpus	100 million	1923-2006
Corpus of American Soap Operas	100 million	2001-2012
British National Corpus (BNC)	100 million	1980s-1993

Source: https://www.english-corpora.org/corpora.asp?b=y[2019-05-08]. Corpus of US Supreme Court Opinions is a dynamic corpus and was last updated in 2017

An online big data corpus may contain millions of or hundreds of millions of words (tokens). In applied linguistics research, it is not possible to analyze the data directly with SEM. But we can extract small data from big data, or extract data which can be analyzed directly with SEM. Section 1.2.3.1 to Section 1.2.3.3 focus on the means of extracting small data from BNC (https://www.english-corpora.org/bnc/).

1.2.3.1 Searching for the frequency of isolated words or symbols

Frequency is the information which can be easily extracted and used in the study of applied linguistics. To search for the frequency, we can employ a range of methods, such as the wildcard, specific symbol and part of speech. When the part of speech is used, we have to be familiar with the correspondence between symbols and meanings (Table 1.2). In this section, we focus on the frequency of isolated words and symbols.

Table 1.2 Part of speech for searching

Variants of symbols				Meanings
[nn*]	NOUN	N	_nn	Common nouns
[np*]	NAME	NP	_np	Proper nouns
[n*]	NOUN+	N+	_n	Common and proper nouns
[vv*]	VERB	V	_vv	Lexical verbs (no do, be, have)
[v*]	VERB+	V+	_v	All verbs (including do, be, have)
[j*]	ADJ	J	_j	Adjectives
[r*]	ADV	R	_r	Adverbs
[p*]	PRON		_p	Pronouns
[i*]	PREP		_i	Prepositions
[a*]	ART		_a	Articles
[d*]	DET		_d	Determiners
[c*]	CONJ		_c	Conjunctions
[x*]	NEG		_x	Negation
[m*]	NUM		_m	Numbers

Source: https://www.english-corpora.org/bnc/[2019-05-08]

Note: In BNC, the web address for any searching result is always https://www.english-corpora.org/bnc/

A. Simple searching for the distribution

If we search for the word "the" in BNC, we can type it into the search query area (Figure 1.4), and click "Find matching strings". The result will show the frequency of the word "the" in the whole corpus.

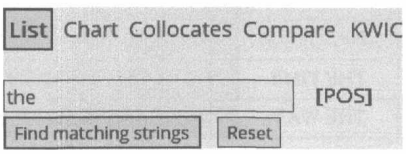

Figure 1.4 Typing "the" into the search query area

Source: https://www.english-corpora.org/bnc/[2019-05-08]

On the home page for searching, if we click "Sections and Chart", we can discover the frequency of the target word in a specific domain covered by the corpus (Figure 1.5). The word "the" appears 409,013 times in spoken English, 834,722 times in fictions, 427,243 times in magazines, 635,001 times in newspapers, 1,136,961 times in non-academic writings,