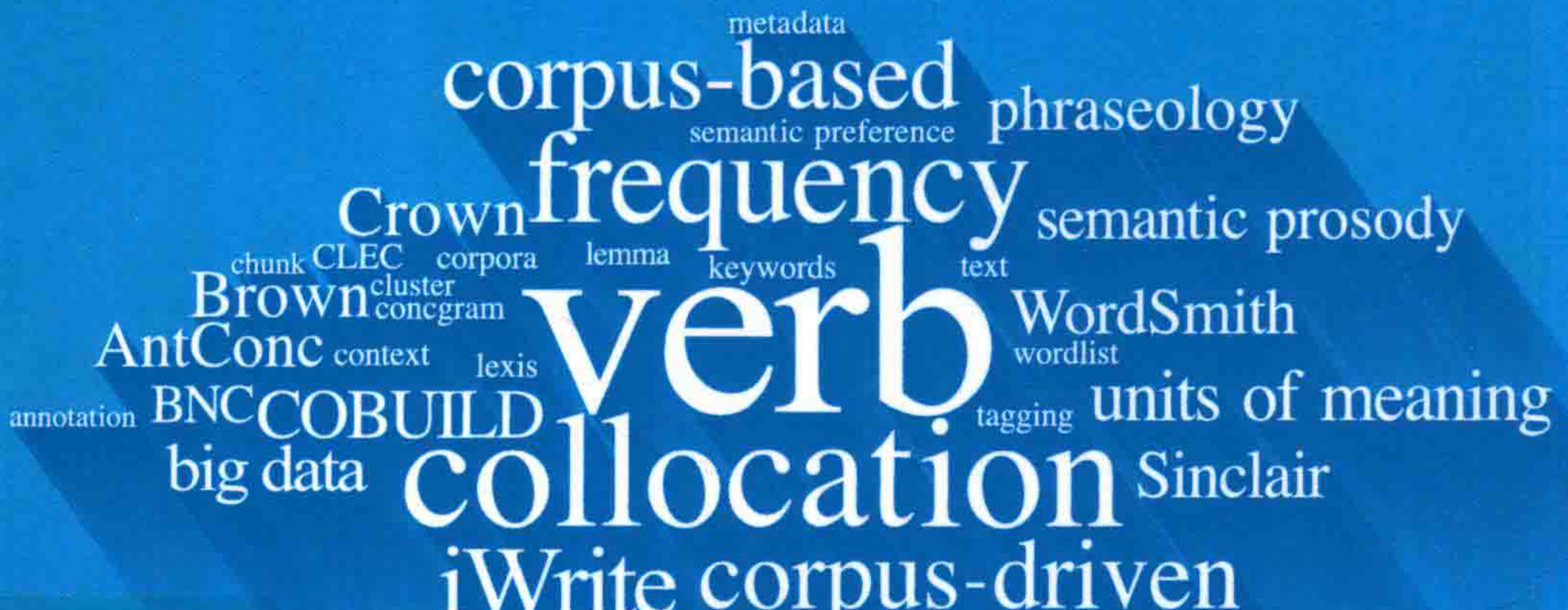
教育部人文社会科学重点研究基地
北京外国语大学中国外语与教育研究中心
大数据视野下的外语与外语学习研究系列丛书

总主编 ◎ **梁茂成**

中国学习者英语书面语动词形式错误自动检查：一项基于链语法的研究

Automatic Detection of Verb Form Errors in Chinese EFL Learners' Written English: A Study Based on Link Grammar

陈 功 ◎ 著



外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

教育部人文社会科学重点研究基地
北京外国语大学中国外语与教育研究中心
大数据视野下的外语与外语学习研究系列丛书
总主编 ◎ 梁茂成

中国学习者英语书面语动词形式错误自动检查：一项基于链语法的研究

Automatic Detection of Verb Form Errors in Chinese EFL Learners' Written English: A Study Based on Link Grammar

陈 功 ◎ 著

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS
北京 BEIJING

图书在版编目 (CIP) 数据

中国学习者英语书面语动词形式错误自动检查 ：一项基于链语法的研究 / 陈功著. -- 北京 ：外语教学与研究出版社，2017.4
（大数据视野下的外语与外语学习研究系列丛书 / 梁茂成主编）
ISBN 978-7-5135-8789-1

Ⅰ. ①中… Ⅱ. ①陈… Ⅲ. ①英语－书面语－动词－研究 Ⅳ. ①H314.2

中国版本图书馆 CIP 数据核字 (2017) 第 099826 号

出 版 人　蔡剑峰
责任编辑　李婉婧
封面设计　彩奇风
出版发行　外语教学与研究出版社
社　　址　北京市西三环北路 19 号（100089）
网　　址　http://www.fltrp.com
印　　刷　北京九州迅驰传媒文化有限公司
开　　本　850×1168　1/16
印　　张　16.75
版　　次　2017 年 5 月第 1 版　2017 年 5 月第 1 次印刷
书　　号　ISBN 978-7-5135-8789-1
定　　价　58.90 元

购书咨询：（010）88819926　电子邮箱：club@fltrp.com
外研书店：https://waiyants.tmall.com
凡印刷、装订质量问题，请联系我社印制部
联系电话：（010）61207896　电子邮箱：zhijian@fltrp.com
凡侵权、盗版书籍线索，请联系我社法律事务部
举报电话：（010）88817519　电子邮箱：banquan@fltrp.com
法律顾问：立方律师事务所　刘旭东律师
　　中咨律师事务所　殷　斌律师
物料号：287890001

本书的研究工作由“教育部人文社会科学研究青年基金项目”资助（项目批准号：14YJC740006）。

本书出版由对外经济贸易大学英语学院学术出版基金资助。

特此鸣谢。

总　序

一、引言

科学研究方法大致有二：其一，归纳法。归纳法指根据一类事物的部分对象的属性推知该类事物的所有对象皆具有某种属性。比如，早期的人类在多次与狼邂逅的过程中，逐渐意识到这种体型匀称协调、四肢修长、头腭尖形、鼻端突出、耳尖直立、善于快速奔跑的野生动物具有极强的攻击性，不可为伍，需要敬而远之或群起而杀之。显然，人类是在经历了多次这样的邂逅之后才意识到了狼的危险性，每一次邂逅都为人类积累了经验、加深了印象，终于在总结若干次教训之后形成了结论：所有的狼都是危险的。诚然，人类在形成结论之前不可能邂逅了所有的狼，但照样可以得出正确的结论。其二，演绎法。演绎法指从一般性的（general）前提出发，通过推导得出具体的（specific）结论。比如，在人们把“所有的狼都是危险的”这一命题视作为一般性前提时，每次邂逅一匹狼，必然会立刻意识到眼前这匹狼是危险的。这其中包含了三个论断，即：所有的狼都是危险的；这是一匹狼；这匹狼是危险的。归纳法是由具体到一般的过程，而演绎法是由一般到具体的过程。

语言研究也不例外，其方法概括起来也不外乎有归纳法和演绎法。演绎法依据可靠的前提进行严密推导，常常可以直击结论。对这种研究方法的运作逻辑我们暂且不做讨论。对于归纳法，其中有若干要素需要考虑。首先，狼有很多特征，哪些特征才具有区别性？哪些属性才是狼的致命属性？比如说，狼嚎是否是我们应该考虑的特征？其次，人类需要与狼邂逅多少次，得出来的结论才是可靠的？返回到语言研究中，前一个问题是语言学家最为关注的问题。语言分析可以从多种语言特征入手，但哪些语言特征才是最有意义的？我们又该如何选择、提取和分析这些语言特征呢？后一个问题是实证研究中的样本问题，即，我们需要

观察多大的语言样本，才可以得出可靠的结论？

自二十世纪后半叶语料库语言学问世以来，研究者越发对自然发生的语言数据产生了依赖，因而产生了“经验主义语言学”、“概率语言学”、“数据驱动语言学”等说法，语料库语言学也随之兴起。就其实质而言，语料库语言学采用的是典型的归纳法。语料库是大量自然语言样本的汇集，解决了以上的第二个问题，即实证研究中的样本问题。有了大样本，归纳而得到结果变得更为可靠甚至可以反复验证。此外，作为方法论的语料库语言学还包含一整套分析方法和分析工具，因而解决了以上的第一个问题，即如何提取和分析语言特征的问题。关于选择何种语言特征进行分析，我们将在下一节讨论。

总之，有了语料库，我们可能“邂逅”的语言事实更为真实、丰富、全面，这也使得通过归纳法得出的结论更为可靠、经得起验证，不需要像Edward Sapir那样亲力亲为地走入印第安部落之中去采集各式各样的语言数据，也不需要像Charles Fries那样随身携带录音机，甚至不需要像Jespersen那样不失时机地以卡片形式随时记录阅读中接触到的各种语言事实。

基于语料库数据进行语言研究，这种方法与演绎法最重要的区别之一在于，研究者在研究中所使用的所有数据均为实际发生的语言事实，而不是靠想象编造出来的句子：

The rat the cat the dog chased killed ate the malt.

Colorless green ideas sleep furiously.

Sincerity admires John.

Golf admires John.

显然，以依据研究者的直觉编造出来的句子作为研究数据，所得结果需要以语言事实来加以验证。正因为语料库语言学研究中的全部数据皆源于事实，结果也更为可靠，因而受到了越来越多研究者的青睐。在这一理念的主导下，我们近年来进行了若干项研究，目的在于利用语料库和语言大数据，对一些语言理论问题进行深入探讨，并试图解决中国外语教育中的一些现实问题。基于这些研究，编辑出版了这一套丛书。

二、语料分析中的语言特征选择

正如狼的所有特征并非同等重要一样，语言特征的选择在语言的量化研究中也至关重要。在前语料库时代，虽有研究者关注语言事实，但

大部分研究者常常根据自己的直觉选择一些特征进行研究。到了语料库时代，特征的选择方法发生了变化。

在语料库时代，人们将语料库中的连续文本制作成词表或多词词表，甚至制作成词类（POS, part of speech）列表或词类序列（POS sequence）列表，然后对基于不同语料库制作而成的此类列表通过精巧的算法进行频率对比，进而有效地发现语料库中更为有意义的语言特征，特别是词语使用方面的特征。这种方法是语料库语言学研究中常用的主题词分析(keywords analysis)，研究中几乎总会使用到一个观察语料库和一个参照语料库，并将由这两个语料库析出的词表进行对比，差异较大的词语(即语言特征）会自动浮现出来。这种特征选择方法虽有人工参与，但研究者的主观性和偏好受到了有效控制，因而研究结果也更为可靠，研究也可以重复验证。在有些研究中，人们还在两个语料库中查询自己感兴趣的语言现象，然后对所得频数进行对比，以发现两语料库间的差异。此外，人们还可以编写复杂的正则表达式，从语料库中提取比词表更复杂的语言特征，如名词短语、介词短语、动宾结构、定中结构、关系从句等。

上文中描述的基于语料库的语言研究是当今最为常见的语言研究方法之一，这种方法的源头至少可以追溯到二十世纪八九十年代，也有研究者将此种研究范式视为盛行于二十世纪五十年代的美国结构主义的延续和发展，甚至也有研究者将语料库之源头追溯到更为久远的时代。笔者认为，基于语料库的研究最早也只能追溯到电子语料库问世之日。正是随着电子语料库的问世，语言研究所需的研究素材在量（quantity）和质（quality）(即语言的真实性）两方面才有了新的突破。基于语料库的语言研究是时代发展的必然，也为语言研究带来了新视野和新维度。在研究过程中，文本的质和量是研究的基础，而文本分析技术和对比算法起到了关键的作用，可以帮助我们发现最有意义的语言特征。

到了当今的大数据时代，情况又有了新的变化。计算机技术的发展推进了网络技术和互联网的普及，而网络的普及就意味着越来越多的人会花费更多的时间浏览越来越多的网页、上传越来越多的内容，发帖、回帖、发表评论，等等，这一切几乎无时无刻不在发生。智能手机的出现和普及更加推进了这一进程，登录网络、发表言论不再受时间和空间的限制。而所有这一切活动中最为常见的媒介正是我们研究的对象——语言。如此发展下去，网络上的语言资源会越来越多，沉淀也会越来越深，长尾效应也越来越明显。在这一背景之下，语言学家自然不应该满足

于原来规模的语料库，他们与计算机领域的专家联手，设计出了各种工具（称之为网络爬虫），可以从网络上获取大量的文本，彻底颠覆了传统语料库的概念。如今，语料库规模已经由原来的百万词级增大到动辄几千万词或数亿词级，甚至达到几十亿或百亿词级。如此规模的语料库，其优势自然毋庸置疑，长尾效应更扩展了研究维度，基于这样的语料库所得到的研究结果也更为可靠、更为多样化，对语言变化的预测能力也更强。然而，在这样的语料库中查询语言特征或由如此规模的语料库生成词语、词类、各类序列或结构列表变得不再那么容易，将这些海量语料库通过主题词分析法进行对比更是不可能了。在大数据时代，我们所面临的问题已经不再是语言研究素材的不足。恰恰相反，数据量过于庞大为语言特征的提取带来了新的挑战，原来的文本分析技术和对比算法不再适用。研究者不得不另辟蹊径。

三、大数据时代的语言研究

大数据给语料库语言学者带来了新问题和新挑战。

数据量（volume）庞大是大数据时代最为显著的特征，但这并不是大数据的唯一特征。数据传输和变化之快，即大数据的速度（velocity）使得研究所依赖的数据几乎没有确定的形态，时刻处于变化之中，体量也不断增大，这也是我们必须面对的另一问题。除此之外，大数据的庞杂性（variety）也是一个棘手的问题。以上三个V被公认是大数据的典型特征。在大数据时代，语料库的创建、语言分析工具、统计分析方法、统计结果的呈现等多个问题都将面临一场革命性的变化。

在语料库创建方面，巨量语料库的提纯是一个至关重要的问题。由于网络文本的多样性，粗暴而盲目地堆砌文本、追求语料库的大容量，会使得语料库变得十分的异质、庞杂，因而是不可取的。为此，人们汲取了网络爬虫技术，并加以改造，推出了Web as Corpus技术并开发了专用软件，依据网络页面中的关键词快速创建各种专题语料库。这种技术必将成为大数据时代语言研究中的重要技术。另外，专题语料库姑且重要，但对于语言研究者而言，语体差异性、文本的时代性等问题也是语言研究中必须考虑的因素。与语体差异性、文本时代性等密切相关的问题之一是，我们应该如何通过各种途径有效获取文本的外部属性（即元信息），这也是大数据时代的语言研究中面临的又一重大挑战。只有挖掘网络文本的元信息特征，研究者才可以利用文本的各种社会属性（如语

种、创作年代、作者身份、作者性别、语体特征、领域特征等），使语言研究特别是文本差异（text variation）研究得以深入。

在语言分析工具方面，由于大量文本都存储于网络或云端，加之语料库规模不断扩大，原先广泛使用的WordSmith Tools、AntConc等单机版的文本分析工具逐渐会变得不再适用，基于网络或云端的工具或许将会成为技术开发的重点之一。此外，在语料库加工方面，基于大数据和深度学习（Deep Learning）技术设计的系统（如谷歌公司开发的句法标注工具SyntaxNet）将代表主流的研究方向，标注的准确率也会有明显提高。

从标注语料库中提取和统计语言特征时，原先广泛使用的统计方法不再适用，主题词分析方法随着语料库规模的增大也必将变得越来越困难，逐渐取而代之的是更为复杂的数据科学（Data Science），聚类、因子分析、复杂回归分析等成为语言分析的常用方法，分析工具也由原来常用的SPSS等工具变成R等更为复杂的系统。R软件的优势不仅在于可以分析大数据，还将编程和统计融合起来，使研究者可以定制各式各样的分析手段。

在统计结果呈现方面，语料库研究常见的图表呈现方式仍然会被广泛使用，但与此同时，随着数据量的增大，数据的可视化将成为呈现研究结果的重要方式，这种呈现方式将更为直观、便于理解。相信在不远的未来，语料库研究的结果将会使越来越多的人受益。

四、结语

随着大数据时代的到来，语料库语言学必将得到更多研究者的重视和青睐，大数据时代的特点将在语言研究中会逐渐显现。我们希望通过本系列丛书的出版推进语言研究的不断科学化，推动我国外语与外语教育研究的发展。

本套丛书是中国外语与教育研究中心“十三五”规划重大项目“大数据视野下的外语与外语学习研究”的研究成果，特此鸣谢。

梁茂成

二零一七年三月

前　言

对自然语言中错误的自动检测是计算语言学研究领域的一个重要课题。已有的英语错误检查系统大多以本族语者为目标用户，只有少数系统专为中国学习者设计；而受各方面因素的制约，学习者错误检查系统的可及性较差，系统查错性能也有待进一步提升。因此，开发一个免费的、以中国英语学习者为目标用户的、查错性能较好的英语错误自动检测工具意义重大。由于时间和精力有限，同时考虑到动词的重要地位，本研究着重探讨中国学习者英语书面语中动词形式错误的自动检查。

本研究采用基于正确规则的方法。具体来说，本研究以型式语法为理论基础，通过链语法这一形式化语法体系，对动词型式语法进行形式化，实现链语法动词词典的重构。本研究还结合中国大学英语四、六级和专业英语四、八级考试大纲要求学生掌握的动词，对链语法词库进行了补充和调整，以更好地实现为中国学习者服务的目的。

本研究动词形式错误检查系统的构建包括四个步骤：（1）动词型式语法形式化准备，包括资源下载、动词提取、文本整理、单词下标设计与添加、屈折形式转换等；（2）大纲动词的处理，包括大纲动词的提取、大纲动词和型式语法动词的查重和型式归类，最后按动词型式将其加入链语法词库；（3）动词型式语法的形式化，包括动词型式的形式化预处理、先导试验、链语法词库的调整、规则编写以及词典重构；（4）重构后词典测试，包括测试语料的确定（学习者错句集和本族者句集各一千句）、预处理和正式测试，最后报告测试的召回率和准确率。

本研究的主要研究结果如下：

（1）本研究所基于的语言学理论（动词型式语法）和形式模型

（链语法）可以较好地适用于中国学生书面语动词形式错误检查系统的构建。（2）本研究对链语法形式化体系的改进，可以有效限制链语法过度的生成能力，提高链接的准确性和分析结果的可识别度。（3）重构后链语法词典的查错性能和句法分析能力得到提高。重构后的链语法词典错句检查的召回率为61.6%，比原词典提高了4.5个百分点；准确率为92.4%，比原词典提高了15.7个百分点。重构后的链语法词典能够分析出949个本族者正确句例，占本族者句集的94.9%，比原词典高出12.2个百分点。

该研究结果也从一个侧面表明，语言学理论成果对于自然语言处理研究有着重要的价值。本研究所基于的型式语法来源于大规模真实语料，是语言学家系统深入观察大规模自然语言得出的语言规则，是对自然语言相对客观的描写。在当前自然语言处理研究领域，基于句法—规则的理性主义方法开始受到质疑，“大规模真实文本的处理成为自然语言处理的主要战略目标”（冯志伟2010a：32）。型式语法正好顺应了这一研究趋势。另外，作为语料库研究成果，型式语法对语言细节描写的充分性和客观性是其他语言学理论所不及的。自然语言处理中形式模型的构建需要这样的语言学知识，因为语言学才是自然语言处理的理论基础，语言学世界里的新发现越多，自然语言处理研究就越能从中受益，“没有明确的语言学知识作为基础的应用领域是走不远的”（冯志伟 2010b；Wintner 2009）。本研究正是在这一思想的指导下做出的初步探索。

在本书的撰写过程中，我的导师梁茂成教授给予了极大的支持和帮助。没有导师的指引和鼓励，本书的写作不可能顺利完成。我还要感谢计算语言学家冯志伟教授、北京外国语大学中国外语与教育研究中心的陈国华教授和李文中教授、北京邮电大学的王小捷教授、对外经济贸易大学的王立非教授，他们在百忙之中抽出时间阅读了本书的初稿，提出了许多宝贵的意见和建议。同时，感谢教育部和笔者所在单位对外经济贸易大学为本研究提供经费支持，在场地、设备等方面予以协助，保证了研究的顺利进行和书稿的最终出版。还有外研社的领导和同事们，他们为本书的出版付出了辛勤的劳动，在此一并谢过。

由于笔者水平所限，书中难免有纰漏之处，恳请各位专家和读者不吝赐教，有不妥之处，敬请批评指正。

目　录

表　目