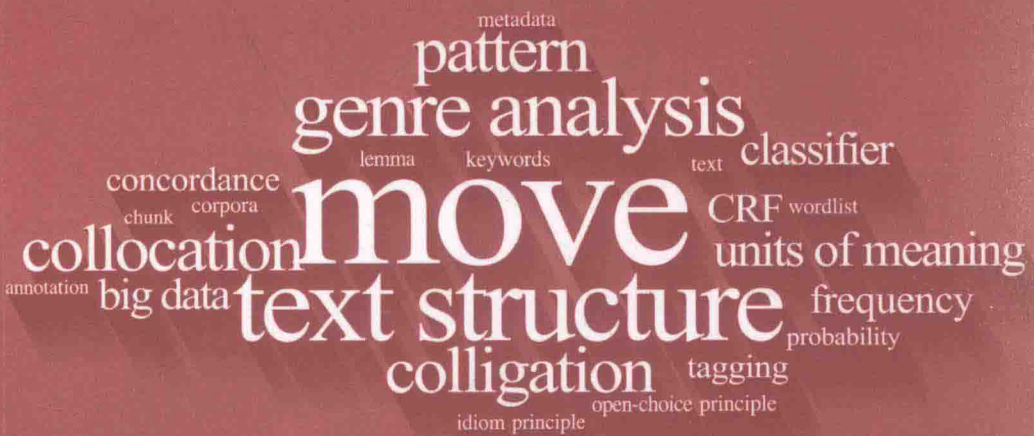
教育部人文社会科学重点研究基地
北京外国语大学中国外语与教育研究中心
大数据视野下的外语与外语学习研究系列丛书
总主编 ◎ 梁茂成

英语学术论文摘要语步结构自动识别模型的构建

刘霞◎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

英语学术论文摘要语步结构自动识别模型的构建

Constructing a Model for the Automatic Identification of Move Structure in English Research Article Abstracts

刘霞◎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

英语学术论文摘要语步结构自动识别模型的构建 / 刘霞著. -- 北京 : 外语教学与研究出版社，2017.11
（大数据视野下的外语与外语学习研究系列丛书 / 梁茂成总主编）
ISBN 978-7-5135-9622-0

Ⅰ. ①英… Ⅱ. ①刘… Ⅲ. ①英语－论文－文摘－语言－系统模型 Ⅳ. ①H31

中国版本图书馆 CIP 数据核字（2017）第 285530 号

出 版 人　徐建忠
责任编辑　毕　争
责任校对　解碧琰　刘　伟
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　650×980　1/16
印　　张　13.5
版　　次　2017 年 11 月第 1 版　2017 年 11 月第 1 次印刷
书　　号　ISBN 978-7-5135-9622-0
定　　价　49.90 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：296220001

该书得到以下项目的资助，特此鸣谢。

项目名称：大数据视野下的外语与外语学习研究

项目类别：教育部人文社科重点研究基地重大研究项目

资助单位：教育部

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20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，充分观察成为可能，归纳而得到的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，我们将在下面讨论。

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Otto Jespersen那样不失时机地以卡片形式随时记录阅读和日常生活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，我们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根本性变化。

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得到了有效控制，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，甚至涉及意义单位。

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，其源头至少可以追溯到20世纪八九十年代，也有研究者将此种研究范式视为盛行于20世纪50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）两方面才有了真正的突破。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（常称为网络爬虫），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，对这些海量语料库通过主题词分析法进行对比则更加困难。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、语言分析工具的开发、统计分析方法的更新和完善、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地异质、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，专题语料库固然重要，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、产生年代、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“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”（编号：17JJD740003）的研究成果，特此鸣谢。

梁茂成

二〇一七年三月

前　言

大数据时代，如何科学、全面地反映一个学科的知识结构和发展状况，一直是学者们关注的焦点。对科学知识形式化的描写，不仅可以帮助新步入一个研究领域的研究者迅速了解领域发展的概况和历史，也可以帮助领域内的专家跟踪和把握学科发展的最新动态，还能够为科技政策的制定者提供比较科学、准确的依据。

挖掘学科知识的需求，直接促使了文献计量学（bibliometrics）的诞生和发展。文献计量学中两种最主要的描述学科领域知识的方法是共引分析（co-citation analysis）和共词分析（co-word analysis）（Besselaar 2006），前者以引文为分析对象，后者以关键词（key words）或索引词（index words）为计量单位。两种分析方法各有利弊，共引分析的结果有时不易解释（Callon 1983），而共词分析容易受作者主观挑选的主题词的影响（Qin 1999），目前常用的知识挖掘软件，如Citespace（Chen 2006，2014），综合了共引分析和共词分析两种方法，并以关键词、标题和摘要共同作为共词分析的研究对象。遗憾的是，由于无法定位摘要的具体的语步类别，即使综合了两种方法和多个研究对象，仍然无法挖掘出学科中最有价值的关键信息，如研究方法和研究结果，因此摘要语步的自动识别研究在这方面具有重要意义。

摘要语步的自动识别模型本质上是一个自动文本分类[1]（automated text categorization）模型。20世纪90年代以来，随着机器学习的兴起，自动文本分类成为信息科学研究的一个重要目标。基于文本分类的研究成果，出现了三类摘要语步自动识别模型，这三类模型各有利弊：1）基于统计构建的词袋模型，虽然能够穷尽词项特征，但对特征不做筛选和归类，导致特征稀疏；2）基于规则提取特征，并对特征进行直觉分类后构建的模型，虽然避免了特征稀疏的问题，但无法全面系统地匹配所有特

1　如无特殊说明，本文提到的文本分类均指自动文本分类。

征；3）结合词袋和语境特征构建的模型，虽然识别效果很好，但只能针对结构化摘要，对大量非结构化摘要的识别效果仍然不能令人满意。针对这种情况，本研究拟结合统计和规则的方法，利用语言学理论提取特征，同时借助语料库语言学、自然语言处理、信息检索技术和统计学等学科中的研究方法，试图构建运行效果更好的能够自动识别常见类型英文摘要语步结构的模型。

本研究模型的构建大体分四个阶段。1）语料的准备和预处理阶段。我们下载了Web of Science数据库收录的《应用语言学》（*Applied Linguistics*）期刊自1993年到2014年出版的所有研究论文摘要，共计440篇，然后清理文本，进行自动词性赋码和句法分析。2）人工标注阶段。由三位相关专业研究人员对语料进行人工标注，标注过程前后持续一年，经历了基于已有研究提出的标注方案自上而下地标注，不带有任何已有的方案自下而上地标注，最后采用了两种方法相结合的方式，并确定了以完整的语句为标注单位的六语步标注方案。经检验，两位标注人员独立标注的一致性较好（Kappa=0.785），最后对独立标注中二者不一致的地方进行多次讨论、修改，达成完全一致。3）提取特征构建模型阶段。人工标注语步结构之后，利用一系列研究工具和方法，提取有效的语步预测特征，再利用这些特征和数据训练学习分类器（learner classifier），获得模型。4）模型的验证阶段。利用构建的模型预测验证集的语步类别，将模型预测的结果与人工标注结果对比，得到模型的识别效果，再与现有的同类模型对比，探索本模型的优势与不足。

本研究的主要发现可以概括为摘要的语步分析、语步结构的有效预测特征和模型的识别效果三方面。第一，本研究突破了传统的语步分析法，基于对大量数据的实际分析印证并完善了已有的语类研究理论。第二，本研究验证了已有模型提取的4个特征的有效性，证实了新加入的3个特征的有效预测力，通过对比发现以语料库的方法提取的新特征比传统方法提取的特征效果更好。从特征的三个维度来看，意义特征对语步的识别度最高（F=0.609），其次是语境特征（F=0.428），识别度最低的是形式特征（F=0.317）。第三，本研究构建了摘要语步结构的自动识别模型，模型的识别效果（F=0.7819）是现有自动识别模型中效果最好的，对信息型摘要的识别效果比现有识别效果最好的模型提高了4.5%。为保证可比性，我们利用同一批语料训练词袋模型AntMover，结果发现本模型比AntMover的识别效果提高了约23%。

摘要语步结构自动识别模型的构建，为下一步学科知识挖掘中定位

摘要的语步和语步内部的关键知识奠定了基础。另一方面，语步的自动识别突破了专门用途英语（English for Specific Purposes，ESP）领域长久以来的人工识别法，为语步分析理论和实证研究走向更多的学科和研究领域，发展成为一个更全面、多视角、多维度的语步分析领域提供了可能。

在本书的撰写过程中，我的导师王立非教授在理论的把握、模型的构建和整个写作过程中给予了我极大的帮助。感谢北外语料库团队的四位导师：许家金教授、梁茂成教授、李文中教授和熊文新教授，他们在我遇到任何困难时，都第一时间给予我指导和帮助。感谢中国外语与教育研究中心的各位老师们，在北外的七年，他们每个人严谨治学的风范令我终身受益。感谢刘振前教授、徐昉教授、苗兴伟教授、梁茂成教授、李文中教授、许家金教授对初稿提出了很多宝贵的修改意见。感谢我的工作单位西南财经大学经贸外语学院的领导和老师对我的支持和帮助。

由于笔者水平有限，书中难免有纰漏之处，恳请各位读者批评指正！

目　录

表 目