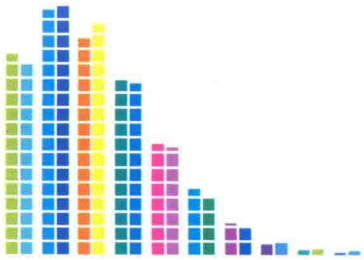
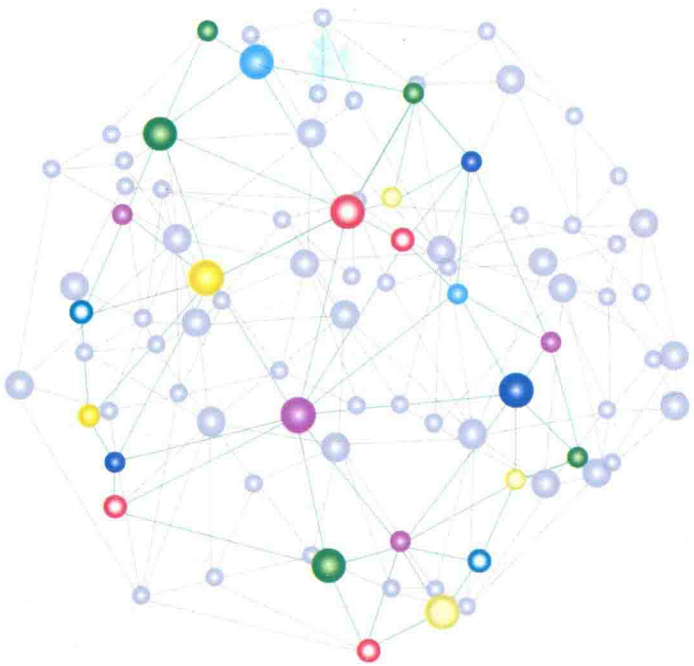
Advances in
Quantitative Linguistics



计量语言学
研究进展

刘海涛◎主编



ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

内容简介

计量语言学是以真实语料为基础，用精确的方法来研究语言结构与发展规律的语言学分支学科。“精确、真实、动态”是计量语言学研究的三个主要特征。本书内容包括：对现代计量语言学基本定律与定律间协同关系的探索，采用语言定律来研究语体或文体的分类问题，采用复杂网络来对语言进行共时与历时的研究，采用计量语言学的方法研究语言规律背后的认知动因，认知约束与语言规律下的语言对比研究，语言能力发展的研究，等等。这些计量语言学研究实践说明，开展以汉语为主要研究对象的计量语言学研究，不但可以更精确地了解汉语的结构特征，而且也可以通过与其他语言的比较，更全面、更深入地理解汉语所具有的特殊性与普遍性，并有助于促进中国语言学的国际化与语言研究的科学化。

ISBN 978-7-308-18075-7
9 787308 180757 >
定价：68.00元

计量语言学

研究进展

刘海涛◎主编

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

图书在版编目（CIP）数据

计量语言学研究进展 / 刘海涛主编. —杭州：浙江大学出版社，2018.10
ISBN 978-7-308-18075-7

Ⅰ.①计… Ⅱ.①刘… Ⅲ.①语言学—计量学—研究
Ⅳ.①H0-05

中国版本图书馆 CIP 数据核字(2018)第 054693 号

计量语言学研究进展

刘海涛　主编

责任编辑　董　唯
责任校对　杨利军　牟杨茜
封面设计　周　灵
出版发行　浙江大学出版社
（杭州市天目山路 148 号　邮政编码 310007）
（网址：http://www.zjupress.com）
排　　版　杭州中大图文设计有限公司
印　　刷　浙江省良渚印刷厂
开　　本　787mm×1092mm　1/16
印　　张　25.25
字　　数　580 千
版 印 次　2018 年 10 月第 1 版　2018 年 10 月第 1 次印刷
书　　号　ISBN 978-7-308-18075-7
定　　价　68.00 元

编委会

◆国家社科基金重大项目“现代汉语计量语言学研究”（11&ZD188）

◆“浙江大学大数据+语言规律与认知创新团队”项目（中央高校基本科研业务费专项资金资助）

* * * * * *

◆国家社科基金西部项目“基于依存句法标注语料的英汉对比研究”（12XYY005）

◆国家社科基金青年项目“现代汉语新闻语体计量研究”（13CYY022）

◆教育部人文社科项目“汉语依存距离的计量与认知研究”（13YJC740112)

◆国家社科基金一般项目“英汉文本特征的计量语言学研究”（15BYY098）

◆国家社科基金重点项目“基于依存句法标注语料库的中国英语学习者句法发展研究”（17AYY021）

◆国家社科基金一般项目“面向自然语言处理的汉语依存句法计量与模拟研究”（17BYY120）

◆国家社科基金一般项目“现代汉语新诗语言计量研究”（17BYY121）

◆国家社科基金一般项目“依存距离视角下的英汉语隐性句法模式研究”（18BYY015）

◆国家社科基金青年项目“基于通用依存树库的依存关系分布跨语言计量研究”（18CYY031）

前　言

计量语言学是以真实语料为基础，用精确的方法来研究语言结构与发展规律的语言学分支学科。这一句话概括了计量语言学研究的三个主要特征：精确、真实、动态。

自2011年年底我们承担的国家社科基金重大项目“现代汉语计量语言学研究”(11&ZD188)立项以来，课题组全体成员同心协力，以现代汉语为主要研究对象，取得了一些有趣的成果，并于2017年1月初免鉴定结项(2017&J004)。至2016年年底时，课题组已在国内外公开发表带有基金批准号的论文80篇。这些论文被以下检索系统收录：SSCI 31篇次，A&HCI 26篇次，SCI 7篇次，CSSCI 20篇次，CSCD 6篇次。两项阶段性研究成果获得教育部第七届高等学校科学研究优秀成果奖(人文社会科学)，三项阶段性成果获得浙江省第十八届哲学社会科学优秀成果奖。在结项后至今不到两年里的时间里，我们又发表了40多篇与本课题密切相关的高水平学术论文。

本书收录18篇文章，大多为首次发表，它们只是课题组成果的一小部分。这些研究涉及的领域包括：对现代计量语言学基本定律与定律间协同关系的探索，采用语言定律来研究语体或文体的分类问题，采用复杂网络来对语言进行共时与历时的研究，采用计量语言学的方法研究语言规律背后的认知动因，认知约束与语言规律下的语言对比研究，语言能力发展的研究，等等。在这里我简要介绍一下这些研究的主要内容。

现代计量语言学有时也被称作齐普夫(Zipf)语言学。这一方面说明齐普夫本人对现代计量语言学的贡献巨大，另一方面也说明其为后人提供了继续前行的路径。齐普夫对于计量语言学最大的贡献是齐普夫定律(Zipf's Law)。该定律指出，文本集中词的出现频率与其序之间存在幂律关系。随后，研究人员发现这种幂律关系普遍存在于自然界和人类社会中。为了解释这一人类社会的普遍现象，科学家提出了很多理论和模型，但却难尽如人意。回到齐普夫定律的本源——人类语言来看，我们至今无法清楚地解释幂律是如何从人类的语言使用中涌现的。实践表明，在语料规模较大时，词频序的幂律关系保持得并不好，而是出现了下弯曲现象，这种下弯曲的语言学意义是什么？句子是人使用语言最基本的单位，词频序在句子中的分布有什么规律吗？于水源等人的“齐普夫定律的语言学解释”首先利用计算机仿真的方法，研究了层级结构可以产生幂律的条件。结果表明，层级结构可以产生幂律。然后，他们又使用英国国家语料库(BNC)和莱比锡(Leipzig)语料库，从词频序关系曲线的下弯曲现象入手，研究了词频序关系随语料规模的增加而改变的原因、词频序在句子中的分布、词频的稳定性等现象和问题。结果显示，随着语料规模的

增加，真实的低频词的频率和新出现词的数量都小于齐普夫曲线拟合的，即它们的增加速度比语料规模增加的速度慢。这不但造成了词频序关系曲线的下弯曲，也使得低频词的频率难以使用一般的统计方法进行研究。词频随语料规模的增加速度显示出了词的不同性质。他们针对各词频序段的词在句子中分布的研究表明，各段词在句子中的分布是均匀的。换言之，齐普夫定律阐述的词频序关系实质上是词在句中的分布规律。这些问题的解决，不但有助于我们搞清楚齐普夫定律的产生及其语言学意义，也有利于我们深入理解现代计量语言学的基本问题。

词长是现代计量语言学中一个长盛不衰的研究热点，这是由于长度容易测量，也便于与人的认知联系在一起。采用拼音文字的语言测量词的长度一般比较简单，要么是按照音节来，要么是按照字母的数量来。对于汉语，特别是现代汉语的书面语来说，由于双音节词占绝对多数，而三、四音节的词又比较少，所以将音节作为测量词的基本单位可能有一定的问题。考虑到汉语书面语的这种情况，我们在对汉语进行词长研究时，可能首先需要解决用什么基本单位来测量词长更合适的问题。陈衡等人在“汉语词长分布计量考察”中，对现代汉语口语和书面语中可能的词长测量单位进行了较详尽的考察，发现在书面汉语的层级结构中，“部件”是词的下一层级；在口语中，“音素”和“音位”都不是词的下一层级，而“音节”是词最可能的下一层级。这个发现也说明，采用拼音文字的语言使用音节多少来测量词的长度是有道理的。以此为测量单位，他们成功地将现代汉语口语和书面语的词长数据与语言单位长度统一分布模型（Zipf-Alekseev Distribution，齐普夫-阿列克谢耶夫分布）进行了拟合，而且发现该模型的参数 a 和 b 是动态关联的，参数 a 的值与语体无关，其直接受到词长测量单位的影响。汉语词长与长度统一模型拟合成功，一方面说明人类语言在长度方面存在普遍性，另一方面也说明不同的语言通往普遍性的具体手段可能会有所不同。语言学是一门探索语言结构模式与演化规律的科学。通过对语言结构模式的探索，我们可以了解到语言的共时面貌，但语言不是一成不变的，只有从历时的角度对语言进行研究之后，才能更好地发现语言的演化规律，才能更清楚地了解为什么语言会有共时的样态，以及更好地预测未来的走向。为此，陈衡等人也对约一千年来汉语词长分布的演化问题进行了初步的历时考察。结果显示，齐普夫-阿列克谢耶夫模型中的参数 a 值随着语龄的增长而变大，增大的原因可能是由于词长增加而导致的。这项研究是至今为止对汉语词长的结构与演化做得最全面的考察，不仅有益于我们对汉语系统的了解，也增强了我们对语言的普遍性与个性、共时与历时关系的了解。

定律（Law）是构成理论的基础，也是现代计量语言学探求的主要目标。采用计量语言学的方式所发现的各种定律虽然有助于人们认识单个语言现象的本质，但如何将这些不同层面的语言学定律结合在一起却不是一件容易的事情。如果我们要对语言做一个全面的描述，那么就有必要将这些定律结合在一起，概括成更高层面的原则，形成一个基于定律的语言学体系或理论架构。1986 年，科勒（Köhler）出版了《语言协同学：词汇的结构以及动力学》（*Zur linguistischen Synergetik：Struktur und Dynamik der Lexik*）一书，这

标志着协同语言学(Synergetic Linguistics)的诞生。协同语言学是将协同学应用到语言研究领域的产物,是计量语言学发展的更高阶段,它的主要任务是提供一套构建语言学理论的架构,即这套建模方法可以用来建立普适的假设,验证假设,并将这些假设组合起来形成定律和类似于定律的描述网络,以便用其来解释所观察到的语言现象。首个协同语言学的模型是由科勒在上述 1986 年出版的德语书中构拟的德语词汇协同子系统。当然,如同任何科学研究领域一样,我们很难用一种语言来说明语言协同系统的普遍性。王璐等人的"汉语词汇结构的协同研究",使用《人民日报》标注语料库,对科勒提出的词汇协同系统模型进行了检验,该模型由多义度、多文度、频率和词长 4 个核心属性组成。结果表明,虽然汉语的结构不同于印欧语言,但词汇协同模型仍是适用于描述汉语词汇结构的。这一研究说明,协同语言学的一般理论与方法具有跨语言的有效性,可能是构拟人类语言系统的一种普适方法。

如前所述,从文本中发现定律是计量语言学家努力的目标。然而,定律本身又具有普适的特征。这自然就引出一个问题,如果许多语言、许多文本都符合一个定律,那么,这定律除了可以反映人类语言的普遍性之外,还有什么用处呢?换言之,我们希望定律在反映普遍性的同时,也可以反映人类语言的多样性。一般来说,一个表示定律的数学公式含有一些参数。这些参数在一定程度上反映了文本或语言的特征。因此,现代计量语言学家在寻求普适定律的同时,也常常采用不同语体(风格)的文本来观察源于定律参数的计量指标是如何反映语言特征的。黄伟等人的"基于词频的现代汉语语体计量研究",采用 15 种词频(谱)计量指标,对现代汉语不同语体的文本进行了统计分析。结果发现,与传统的型例比(TTR)指标相比,基于词频的文本计量指标不会受文本长度的影响,其中与实词、虚词使用频率有关的指标 indicator-a、表示词频谱集中程度的指标 $RR_{s,rel}$ 和 $RR_{s,rel2}$、表示非常用词覆盖率的指标 R_3 均具有区别语体的作用;indicator-a 和 $RR_{s,rel2}$ 这两个指标不受文本长度的影响,能够很好地区分口语体与书面语体的文本。这一研究将语言学定律蕴含的普遍性与特殊性联系在了一起,对于语言学家更深入地了解定律在语言研究中的作用具有很大的价值。

语言是人类最重要的交流工具,而交流在很大程度上指的是信息的交流。熵是测量语言所含信息量的一个指标,一般用来表示语言特征携带的平均信息量或其频率分布的均匀或丰富程度。熵值越高,表明语言特征的分布越均匀,用法越丰富;反之,熵值越低,表明语言特征的分布越不均匀,用法越固定。陈蕊娜等人的"现代汉语'熵'的语体差异",从兰卡斯特现代汉语语料库(LCMC)中选出 5 个代表性语体,研究比较了各语体中句子不同位置上的词和词性及一元词和二元词的熵值差异。结果表明,词的相对熵在新闻中最高,在散文、公文、学术和小说中依次递减;词性相对熵在小说中最高,在散文、新闻、学术和公文中依次递减。词性相对熵作为衡量不同语体句法灵活度可能更为可靠,统计检验表明词性相对熵在不同语体间,尤其是在"叙述"和"说明"语体间差异显著。一元词除形容词的熵值在散文和小说中差异不显著外,其余名词、动词和数词在 5 种语体中差异均显

著。除某些特殊的词性组合外，二元词的熵值在大多数语体间比较均差异显著。陈蕊娜等人的研究验证了对语体的某些直觉理解，如叙述类语体携带信息量较大，而说明性语体信息量较少，这一研究也将计量语言学的研究从形式扩展到了内容。

从形式走向内容的另一个大的领域是文学。采用计量的方法来研究文学家关心的问题已有很长的历史了。与文学相关的语言计量研究，在一定程度上，可能要早于纯粹的计量语言学研究。就汉语当代文学而言，新诗与散文的关系一直是许多人关心的问题。为了寻求新诗的文体特征，潘夏星等人的“汉语新诗与散文的文体计量研究”从文本的词语入手，选用高频词、词汇丰富度等计量指标，对不同文本进行了分析。结果显示，散文高频词的描述对象呈现出多样化的特征，而新诗的高频词则表现出较明显的时代特性；散文文本的词汇丰富度明显低于新诗文本；词语频率的分布特征显示，散文文本表现出较强的“自组织性”，大部分新诗表现出了一定的“自组织性”，但也有部分新诗体现出了诗人较强的人工参与度。词类频次的分析结果则显示，散文文本安排各种词类的方式明显与新诗文本不同。两类文本在名词、代词的使用频度上表现出一定的相似性，这可能是新诗“散文性”的根源。这一研究从一定程度上解决了新诗的文体定位问题，展现了用计量方法研究文学问题的适用性和可行性。

语言定律的发现及应用是现代计量语言学的主要任务。通过定律以及定律之间的协同关系，我们可以在一定程度上解决语言作为一个自适应系统的一些问题。然而，语言不是一个简单的适应系统，而是一种人驱复杂的适应系统。这里说的复杂主要指的是，语言系统像其他许多系统一样也存在“部分之和不等于整体”的系统特征，存在着涌现的现象。近年来，为了寻求语言结构局部与整体之间的关系，语言学家也开始采用复杂网络的方法对人类语言的诸多方面进行了探究。对于语言学家而言，“复杂网络是工具，而不是目的”，这就要求我们应该以问题为导向，采用复杂网络方法来解决一些用传统手段不易处理的问题。汉语词的形成机制可能就属于这类问题。我们知道，在现代汉语中，二字词居多，但任何汉语书面语文本都是由字组成的，词与词之间并没有明显的界限。那么，汉语的词是如何从这些连续的字串中涌现出来的呢？丛进等人的“基于网络方法的现代汉语二字词形成机制研究”采用复杂网络方法研究了这个有趣的问题。作者首先假设，如果一个字同现对的频数大于这两个字与其他字形成的同现对的频数，则该同现对就是一个二字词。他们基于兰卡斯特现代汉语语料库的两个子库，构建了两个有向有权字同现网络，在这种网络里满足上述条件的字同现对也可称为二节点边岛屿。结果发现，从网络中提取到的二节点边岛屿基本上都是二字词。而且，所有的二字词都有机会形成二节点边岛屿。这一研究为汉语二字词作为结构性的语言单位如何在其局部语境中突显出来提供了明确的机制，是我们理解语言涌现的一个好例子。

复杂网络不但可以对语言结构模式的涌现进行共时的研究，也可以从整体的角度来探索语言的演化规律。陈芯莹等人的“语言网络中的汉语单字词演化”选用 4 个不同时期的汉语真实文本分别构建了上古汉语、中古汉语、近代汉语和现代汉语的字同现网络，并

对这些网络的整体特征以及“在”和“人”两个单字词在不同时期语言网络中的特征变化进行了对比分析。研究结果显示，利用新兴的网络分析方法和工具，可以更容易、更直观地观察到汉语系统整体和个别语言现象的历时演变趋势，同时为这些演变趋势提供量化数据的支持。该文展示了如何利用网络分析方法获取传统研究方法难以测量的演化区别特征，实现了对语言系统和个别语言现象的共演分析，展示了网络分析方法在语言演化研究中的巨大潜力。

现代语言学认为，语言是一个符号系统。按照一般的逻辑，既然语言是一个符号系统，我们当然有理由采用各种研究符号的方法来研究语言。既然语言是一个符号系统，我们当然也可以对它进行全方位的数理剖析，也当然有权利采用常人一辈子也不说的语句来探求所谓的语言规则。然而，经过几代人的努力，我们发现这些抽象的成果似乎离现实中的人类语言越来越远。如果语言学是研究语言结构与演化的规律，而语言学家找到的规律却不是我们日常使用的语言的规律，我们要这样的规律干什么呢？这一问题的根源可能在于，当语言学家越来越痴迷于生活在自己构拟的符号与形式化的迷宫时，他们也就远离了人类正常的语言，离现实中的语言越来越远。在这种情况下，我们怎么能期待他们所寻找的规律还是人类语言的规律？这可能也是人类语言学家研究人类语言的成果反而更适合描写机器使用的形式语言的一个原因。为了摆脱这种“只见语言，不见人”的窘境，我们有理由将语言视为一种由人驱动的符号系统。换言之，人类语言规律可能只是人类认知规律的一种反映。我们不能无视人类认知的约束与限制，抽象地谈人类语言的某种超人的属性。否则，这些讨论即使再高深，再抽象，可能也不是人类语言的规律，而只是脱离现实的数学演算。在科勒、王璐等人有关德语与汉语词汇的协同模型中，人类语言系统得以运作的根本就取决于人类由于受认知机制所限而出现的语言使用的省力原则。这种省力原则对于语言的约束，当然不只限于词汇子系统。

依存距离最小化是近年来通过大规模多语种真实语料发现的人类语言的一个普遍特征。它说的是，人在造句的时候，更倾向于选择句中词语的某些线性排列，在这样的线性排列中，具有句法关系的词语之间的线性距离（依存距离）之和具有最小化的倾向。依存距离最小化可能是人类进行语言处理所遵循的省力原则在句法层面的体现。值得注意的是，尽管省力原则可能是许多语言特征或规律形成的主要动因，但不同的语言，或者在同一语言的不同层面，人们使用省力的手段可能会有所不同。这不但体现了语言的多样性，而且也反映了语言作为一种复杂系统应该具有的适应能力。陆前等人的“交叉、根位置与组块对依存距离的影响”采用计算机仿真依存结构树的方法，对交叉依存、根节点位置和组块等可能导致依存距离最小化的因素进行了详细分析。通过与真实语料的统计数据进行对比，结果发现，上述三者在一定程度上均能减小依存距离，并在句子线性组配中起重要的作用。这一研究还发现，为了降低长句子的依存距离，人类会构拟一种比词大的动态语言单位（组块）来减小依存距离。这说明，语言作为一种复杂适应系统，当需要在有限的认知资源约束下处理某些按照常规方式难以处理的任务时，会自动采取某些临时性的处

理方法以便获得有限资源下的问题求解方案。

从形式上说，长距离依存关系是造成句子难以理解的根本原因之一。既然如此，为什么人们还要使用这些看似不省力的长距离依存关系呢？研究表明，依存距离的概率分布基本符合幂律，而语言相关幂律一般反映了交际双方省力达到平衡的一种状态。具体来说，就依存距离而言，在真实的语言中，我们可能也需要平衡最省力与表达精确之间的矛盾。回到长距离依存关系，我们的具体问题是：真实语言中使用的长距离依存关系是否一定会增加语言处理的难度？这些长距离句法关系中是否存在其他因素在某种程度上消减了依存距离的影响？这两个问题也是徐春山等人的“现代汉语介词‘在’与主语的依存距离研究”的主题。为了回答这两个问题，他们首先对汉语中高频介词“在”的依存距离做了阅读实验，结果表明，就涉及介词“在”的句法关系而言，依存距离的变化并不会显著地影响其句法处理的速度。这说明，依存距离并不一定会增加句法处理的复杂度。其原因可能与介词“在”独特的句法语义功能以及高词频等特征有关。介词在语言中一般充当结构标记，往往涉及较长的依存距离。因此，介词对依存距离不敏感这个特点是符合语言省力原则的。此外，对主谓关系的研究发现，新主语的依存距离普遍较短，已经出现过的主语的依存距离则普遍较长，而零（省略）主语的依存距离最长。这说明，依存距离可能与主语的信息特征有关，熟悉程度较高的主语对依存距离可能不敏感，反之亦然。句子的信息结构一般为旧信息在前，新信息在后，而主语大都位于句首，一般为旧信息，因此对依存距离可能也不敏感，这也符合语言的省力原则。这一研究说明，就语言系统而言，许多因素是交织在一起的，我们很难通过测量一个因素或特征来了解一个如此复杂的系统的运作机理。然而，尽管这些反映系统各个方面的特征会有所不同，但统领这些不同的认知机制是相同的。换言之，语言研究的根本任务可能就是发现在普遍认知规律下如此多样的人类语言是如何运作的，这些多种多样的语言规律是如何在受到普遍认知规律约束的情况下，又可以满足具体交际需要的。

从动态或语言处理的角度看，依存距离反映了人在处理句子时的认知压力。从静态的角度看，依存距离反映了句子结构的共时复杂性。按照历时的语言结构可能随着时间的变化而变化的语言演化模式，如果有适当的历时句法标注语料库，我们有可能据此来探究语言句法结构复杂程度的演化路径。如果我们将业已证明可反映语言语序类型特征的依存方向（构成依存关系的两个词之间的相对位置）作为一种计量指标来一起使用的话，也可以观察语言语序类型的历时演化情况。刘丙丽等人的“汉语白话依存距离与依存方向历时统计分析”，通过自建的前期、唐五代、宋代、元明、清代、现代等句法标注语料库，对不同时期文本的依存距离与依存方向进行了统计与分析。结果表明，汉语语序具有支配词居后的历时演变倾向，而且句子的平均依存距离是持续增大的，这意味着汉语的句法结构有复杂化的倾向。从依存方向的角度看，汉语的主要语序似乎并没有发生显著性的变化。这一研究也提出了一个值得进一步思考的假设，如果随着表达精确性或所表达内容复杂化的需要，汉语的句子结构也会变得更复杂，以便满足这些需要。这是符合逻辑的，

也是可以理解的。有趣的地方可能在于，依存距离不但反映了句子的复杂性，而且也与认知难度有关，这样一来，是否也意味着从古到今，讲汉语的人的认知压力一直在增加呢？如果是，为什么人们要增加句子的复杂程度，从而使自己交流起来更费力呢？现实可能是，人们并没有感觉到这种压力，这种无感难道是由于人的相关认知机制也随语言的变化而变化所产生的吗？这是一个交织了语言与认知共演的有趣问题，值得深入研究。

语言的普遍性与特殊性一般需要通过对多种语言进行细致的对比后，才能得到更可靠的认识。依存距离与依存方向作为两个基于句法标注语料库的计量指标也为双语或多语句法特征的对比研究提供了客观的分析手段。李雯雯等人的“汉英主宾语句法计量特征的对比研究”采用自建的汉英双语依存树库，对汉英主语和宾语依存关系的计量特征进行了统计分析。研究显示，汉语主语和宾语依存关系的依存距离均值均高于英语，这可能说明汉语比英语承载了更多的认知成本和工作记忆负荷；不仅在一种语言当中，在具体的依存关系中，如主语和宾语关系中，依存距离也存在最小化倾向；从语序类型学的角度看，汉语和英语都是典型的 SV 及 VO 语言。这项研究再次验证了不同的语言可能会有不同的依存距离，从而再次引发了关于语言与思维关系的思考。接着前面在介绍刘丙丽、徐春山等人的研究时引出的问题说，如果一种语言的依存距离总是大于某些语言，而讲这种语言的人又没有感觉到明显的认知压力，那么是这种语言改善了讲这种语言的人的认知能力，还是这种语言的结构中有一些虽然可以增加依存距离，但又不会带来认知压力的语言成分呢？抑或是不同的语言会启动不同的复杂适应机制来降低人们在处理语言时的认知压力呢？这一问题的解决，可能需要多个学科学者的努力，但计量方法毫无疑问有助于我们对这一复杂问题的认识以及解决方法的寻求。

词类分布是人类语言中一个很重要的不变量，在所有词类中，名词所占比例是最大的，约占到 40%，而且这种不变性具有跨语言的普适性。作为真实语言中使用最多的词类，名词在具体语言中除了“光杆”名词外，更多的是以短语形式出现的。这些长短不一的名词短语，在数量上，绝对是构成句子的主要力量。从计量语言学的角度看，探求语言单位长度相关的规律一直是学者们的研究重点之一。短语长度的研究，由于资源的缺乏，一直鲜有人做。王华等人的“汉英名词短语长度的计量研究”采用宾州中文树库、宾州英语树库和国际英语语料库的英国英语部分，研究了汉语和英语中名词短语长度分布的计量特征。文章研究了名词短语的基本量化特征，并用可以描述多种语言的语言单位统一长度模型对汉英名词短语长度的分布进行了拟合检验。结果显示，该模型可以用来描述汉语和英语中的名词短语长度分布，并初步发现参数 b 可以体现一定的语际差异。这一研究一方面可能说明统一或普适的语言单位长度分布可能是人类某种共同的认知机制约束的结果。另一方面，这项研究也有助于我们了解人类语言中占比最大的词类在真实语言中的结构模式分布情况。

语码转换是现代语言生活中一种常见的现象。但对于这一问题的研究，从社会语言学角度出发的比较多，从句法，特别是采用计量方法探求语码转换规律的研究则非常少。

王琳等人的“汉英语码转换的句法计量分析”基于自建的依存句法树库，对汉英语码转换的句法进行了初步的计量分析。研究发现，汉英语码转换的词类、动词所支配和充当的主要句法关系均符合齐普夫-阿列克谢耶夫分布。构成汉英语码转换的主要词类是动词、名词、形容词、数词、副词等，动词是汉英语码转换句子结构的中心，动词所支配的主要句法关系有状语、宾语、补语、主语等；主语、定语和状语关系以支配词居后的依存关系为主，宾语关系以支配词居前的依存关系为主。这一研究不仅向我们展现了语码转换这种语言混杂现象也是有规律可循的，而且他们发现了含有混合句法关系的依存距离要比单语的大，说明语码转换现象可能也是某些认知机制约束的结果。

语言发展或习得规律的探寻不仅有助于语言教学方法的改进，也有助于了解人的认知是如何与语言能力共同进步的。蒋景阳与刘舜佳的“高中学习者英语写作词汇运用发展的计量研究”以三组不同水平的高中英语学习者的限时作文为语料，利用多种文本分析工具，从词汇多样性、词汇密度、词汇复杂性和词汇频率分布四个维度，对不同写作阶段词汇运用的发展模式和特点进行了定量研究。研究发现，在作文长度恒定时，词汇多样性随写作水平提高呈现出非线性的发展趋势，尤其是在高水平组出现了“词汇高原”现象；排除重复实词后，统计得到的词汇密度从低水平组到高水平组有增长的趋势，但仍存在一定的弹性；在词汇复杂性方面，从低水平组到高水平组，高中学习者对最常用词的使用逐渐减少，而在对次常用词和非常用词的使用上则呈非线性增长，总体呈现“泰迪熊”现象；在词汇频率分布方面，本研究利用齐普夫定律验证了学习者写作文本的规律性，同时也说明高频词在语言学习中的重要性。这一研究虽然没有使用更复杂的计量指标，但已经在一定程度上展现了计量方法在语言习得与语言发展领域的应用潜力。

句法习得是语言习得研究或语言能力发展中的一个重要环节。一般来说，母语主要句法关系的习得在 2 岁左右时就基本达到了成人的水平。而在这个年龄段，儿童的一般认知水平尚未成熟，可以用语言所表达的事物也比较简单，这使得我们在研究儿童句法习得时可能会遇到认知能力不足造成的这样或那样的问题与困难。聋人语言是语言研究的一种宝贵资源。聋人大多由于自幼失聪，不能像健听人一样依靠听觉获得正常的语言输入，因而大大延缓了句法能力的发展过程；当学龄聋人进入学校时，虽然他们汉语的句法水平可能达不到 2 岁健听儿童的水平，但其认知能力或大脑中的概念网络却可能与同龄的健听儿童差别不大。在这种情况下，研究聋生的句法发展，有可能使我们更清楚地掌握句法系统的发展或形成规律。金慧媛等人的“基于依存语法的聋生汉语书面语句法能力发展研究”创建了跨度为 9 个年级的聋生汉语书面语依存句法树库，通过依存距离、依存方向以及依存关系的构成等诸多方面，折射出聋人汉语书面语的句法复杂度及其句法能力的发展等问题。研究表明，随着年级的不断升高，聋生汉语书面语的句法能力显著提高；与健听人相比，接近成年的聋生汉语书面语的句法能力与健听人仍存在较大的差距。他们也对小学、初中、高中三个阶段的聋生汉语书面语的依存距离进行了分析，结果发现，其依存距离的概率分布均符合齐普夫-阿列克谢耶夫模型。

正如我们此前所说的那样，与同龄健听儿童相比，聋生的概念系统是差不多的，弱的是句法系统。对于汉语这样的孤立语来说，词语在句中的句法功能主要是通过语序以及虚词来实现的。因此，虚词在研究聋人句法发展的过程中，可能起着重要的作用。严菁琦等人的“基于树库的聋生书面语介词句法发展计量研究”采用依存树库，通过依存距离以及分布规律，研究了小学到高中阶段的聋生对介词相关的依存关系的句法发展和句法水平特点。结果发现，聋生的介词使用在不同阶段有不同的特点，低年级和中年级阶段，介词平均依存距离发展变动大，但阶段内变化趋势不明显。高年级第三年平均依存距离呈稳定而显著的增长。介宾结构和状语的不同依存距离分布的数据显示，聋生早期优先使用介词句法关系中的短距离相邻结构，平均依存距离低于健听人的口语水平，之后逐渐增加介词的其他用法，以及长距离结构的使用，到了高年级依存距离接近健听人。对观察到的依存关系和依存方向的异常值进行具体分析后发现，介词缺失、语序不当等因素可能也是导致平均依存距离变短的原因之一，这进而使得聋生介词的依存距离在初期阶段始终低于健听人的语言水平。研究也发现，聋生在介词相关句法关系依存距离的变化可能也受到了聋人认知能力发展及手语句法的影响。这与前一项有关聋生句法的研究也有助于我们了解从概念网络映射到线性序列的过程中，句法是如何起作用的，以及句法系统是如何在使用中形成的。

以上这些研究的对象主要是汉语，然而，计量语言学方法的科学性使其很容易地可用于研究其他语言。对于计量语言学而言，有关新语言的研究是至关重要的，新语言不但可以证实或证伪已有的语言定律与理论，也有益于发现新的定律与方法。那日松等人的“蒙古语词长的计量研究”考察了语料规模、文本大小、不同文体对蒙古语词长分布的影响，得到了适合于描述蒙古语词长频率分布的 4 类分布，这一结论与其他语言词长频率分布结果是一致的，进一步证实了不同语言中词长频率分布的共性。然而，虽然存在共性，但是在不同规模语料中蒙古语词长的频率分布也有自己的特点：静态语料（词典）中蒙古语词长频率分布更倾向于服从康威-麦克斯韦-泊松分布（Conway-Maxwell-Poisson）；同分类语料和独立文本中蒙古语词长频率分布更适合于用扩展正二项式分布（Extended positive binomial）来描述；在“小说”“散文”“诗歌”文体中更容易发现蒙古语词长的频率分布规律。在不同文本大小、不同文体中的蒙古语词长与词频关系研究中，他们证实了 2000 词左右的独立文本中存在蒙古语词长与词频的幂律关系，而且在“长文本”“小说”“散文”“诗歌”中词长与词频的幂函数拟合结果更好。那日松等人的研究扩展了在不同边界条件下的（蒙古语）词长研究，比较系统和全面地探索了蒙古语词长分布、蒙古语词长与词频的关系。这项研究的意义，不仅在于我们对蒙古语的词长有了更科学的认识，还在于扩充了人类在词长方面的知识库，使我们对词长相关的人类语言的计量特征有了更深入的了解。

以上这些研究不仅加深了我们对于汉语的认识，也扩充了计量语言学研究的理论与应用领域。本书是集体合作的产物。我作为课题的首席专家，几乎全程参与了以上 18 篇文章的研究与写作工作。统稿期间，梁君英、王亚蓝、林燕妮、方昱、牛若晨做了大量的工

作。我们感谢课题组所有成员在过去五年间的合作，没有大家的共同努力，我们很难在短时间内取得如此之多的成果。全体作者一致同意将本书献给冯志伟先生八十华诞，以感谢他长久以来对我们的支持与帮助，感谢他为中国语言学研究科学化而做出的不懈努力。

我们的计量语言学研究实践说明，开展以汉语为主要研究对象的计量语言学研究，不但可以更精确地了解汉语的结构特征，而且也可以通过与其他语言的比较，更深入地理解汉语所具有的特殊性与普遍性。汉语计量语言学研究是对国际计量语言学的全面补充与发展，有助于提高中国语言学界在国际学界的声望与话语权。来吧，让我们在语言研究科学化的道路上同行。

刘海涛

2018 年 9 月于启真湖畔

http://person.zju.edu.cn/lht

目　录