大数据时代的汉语语言学研究

詹卫东
北京大学中文系/中国语言学研究中心/计算语言学教育部重点实验室

意大利哲人尼可罗·马基亚维利（Niccolò Machiavelli）有句名言：“风景画家要描绘山峦之美，需先置身卑谷”。要思考今天这个时代如何去做语言学的研究，恐怕也应该跳出语言学自身的藩篱，放眼环顾我们身处的社会环境和学术生态，从时代进步的节奏和邻近相关学科的发展脉络中，或可反观语言学的律动轨迹，把握未来的方向。

一 身处大数据时代

近几年来，“大数据”（Big Data）这个词见诸媒体越来越频繁，无论是企业界，还是学术界，“大数据”都是一个正在迅速升温的热门话题。2013 年年初，反映这一时代特征的代表性著作，舍恩伯格和库克耶合著的《大数据时代：生活、工作、与思维的大变革》中文版和英文版先后面世[1]，为“2013 年是大数据元年”提供了最好的注脚。正如该书副标题所宣称的，大数据是一场大变革，从生活到工作，乃至思维方式，影响可谓既广又深。书中给出了很多佐证这一观点的例子，这些令人印象深刻的例子，其引人入胜之处，既在于它们直接解决了大众生活中面临的一些普遍问题，同时又在于其解决之道正是引入了大规模数据资源和先进的数据分析技术。这里举其中两例略加说明，一个例子是商业消费领域的：研究人员从网上飞机票预定系统的机票销售历史数据中，提取机票价格随时间波动的趋势性规律，从而为人们选择恰当的购票时机，节省购买成本（Etzioni et. al, 2003）[2]。这个非常典型的基于大数据的商业应用系统，无疑对于企业，对于普通消费者都有很强的吸引力，通过大数据计算，直接为社会创造了经济价值。另一个例子是公共卫生领域的：研究人员发现，从人们在 Google 搜索引擎中输入的最常见的 5000 万个检索关键词数据中，可以找到一些特定的检索模式。这些模式跟美国疾控中心存储的季节性流感爆发期间的病例报告数据之间有很高的相关性，从而可以依据这些检索摸式，加上分布在全美各地的以百万计的 Google 用户的实时查询数据，来估计季节性流感在美国各州的传播态势。传统的基于实际病例统计和实验室病毒分析的方法通常需要一到两周时间才能给出报告，而基于用户查询记录的大数据方法则可以做到每天都发布流感态势报告[3]。这项研究成果发表在 2009 年 2 月份的《自然》（Nature）杂志上（Ginsberg et.al, 2009）。

大数据处理的魅力不仅体现在上述典型的数据计算和分析领域，而且也开始在一些人文社会科学研究领域初试啼声。比如基于维基百科（Wikipedia）[4]的大规模文本分析来从某些特定角度展示人类历史变迁的宏观面貌，就是一个有代表性的例子(Leetaru, 2012)。研究人员利用一台有 4000 个 CPU，内存为 64TB（1TB=1000GB，即 10^{12} 字节）的超级计算机对 400 万篇以英语写的百科文章进行处理，提取其中的日期、地点信息，并通过统计每篇文章中的情感词，计算某个特定年份的情感指数（取值从极负面到极正面分为 200 级），用这种方法计算了 1800-2012 年间 212 年的情感指数，并将计算结果用 212 张叠加的世界地图来表示，

---

[1] 中文翻译版先于英文原著出版，也可以算是信息时代的一个有趣特点。

[2] 这项研究后来得到风险投资成立了名为 Farecast 的公司，该公司后来又被微软公司以 1.1 亿美元收购，集成到微软 Bing 搜索引擎中为用户提供服务（http://www.bing.com/travel）。

[3] Google 流感趋势网站（http://www.google.org/flutrends/us/#US）将流感状态分为“极轻、轻、中等、重、极重”五级，并以五种颜色区分，在 Google 地图上标记每个地区的流感状态。

[4] http://www.wikipedia.org/ 有 285 种语言，超过 2200 万篇文章。

即以地图上不同地点标记不同颜色来反应当地在某个特定年份的情感指数。这 212 张按年叠加带有颜色变化的世界地图以可视化（Visualization）的方式在网页上呈现[5]，在某种程度上，可以说是直接把一幅“风云际会、斗转星移”的世界史动态画卷铺展在了读者面前[6]。

毫无疑问，“大数据”已经给这个时代打下了鲜明的印记。身处其中，无论是否愿意，都将受其影响。就科学研究而言，有的学科因为直接涉及大数据分析已经发生显著的变化，比如计算机科学与语言学的交叉学科——计算语言学在近十年来的发展，就是这样的例证。下面不妨快速扫描一下这门年轻的学科近半个世纪的发展历程，可以更深刻地体会“大数据”对一个学科所带来的冲击和影响。汉语语言学未来的进程或可从中有所借鉴。

二 来自计算语言学的启示

计算语言学（Computational Linguistics）从其应用目标来说，就是让计算机能够“理解”人类的自然语言（Natural Language Understanding，NLU）。这个任务的实质是希望找到从语言的形式映射到语言的意义的机械方法。如果把“理解”人类的自然语言看作是人类智能行为的主要特征，那么，自然语言理解显然属于人工智能的研究范畴，即探求作为高级智能特征的人的语言行为在多大程度上可以机械化。

作为一个仅仅是诞生在半个世纪前的相当新的研究领域，NLU 所经历的发展过程称得上是跌宕起伏。而伴随其间的，可以说正是在 NLU 不同发展阶段人们对于其处理对象——“自然语言”的不同理解。众所周知，NLU 脱胎于机器翻译（Machine Translation，MT）。上世纪中叶，自动计算装置在二战中破译密码的威力在战后继续发酵，刺激着正在重建新世界的人们的想象力。用刚问世不久的电子计算机把一种自然语言翻译成另一种自然语言顺理成章地也被看作是密码破译的过程。MT 从想法到能实际运行的演示系统，只用了短短五年时间（1949-1954）。然而，由美国 Georgetown 大学和 IBM 联合研制的第一个 MT 系统只是在媒体宣传和争取政府资助上获得了实实在在的成功，真正能够服务于社会解决翻译问题的 MT 系统并没有像其研制者所宣称的那样在三五年内就问世。相反，1966 年发布的著名的 ALPAC[7] 报告终结了 MT 的头一个十年热潮。人们开始透过计算机难以逾越的翻译障碍反思人类自然语言本身的性质。翻译不再仅仅被简单地看做是密码破译的信息处理过程，自然语言也不仅仅是一串单词的序列。如何让计算机真正“理解”人类的自然语言，语言的层次结构该如何分析、如何从形式结构映射到意义表示，等等理论问题开始引起来自计算机科学、数学、语言学等跨学科研究人员的深思，新兴的交叉学科——计算语言学也正是在这个背景中走上历史舞台的。上世纪七八十年代，各种分析自然语言的形式理论和计算方法如雨后春笋般涌现，其中著名的如基于概念依存图（Concept Dependency Graph）的知识表示方法与脚本理论（Script Theory），广义短语结构语法（GPSG），词汇功能语法（LFG），中心词驱动的短语结构语法（HPSG），扩充的递归转移网络（ATN）等等，均各领一时风骚，不仅如此，语言学家提出的一些并不是直接要为计算机服务的语言学理论（如系统功能语法）也受到计算机科研人员的重视，被用来作为计算机模拟人类语言行为的理论指导（冯志伟，2010）。在经历了 MT 被工业界和政府科研经费资助冷落十年之后，科学家们在“理解”人类自然语言方面所取得的进展为 NLU 燃起了希望。这个阶段，NLU 躺在语言学的怀抱里，自然语言在理性主义哲学的关照下被看作是有限结构（有限规则）的无限递归应用，MT 的主流是基于规则的方法，计算机程序首先按照语言学理论提供的规则模型，拆解原文的结构，把原文句子分

[5] https://www.youtube.com/watch?v=KmCQVIVpzWg

[6] 需要说明的是，尽管上述这些大数据计算的例子确有震撼效果，但客观而言，大数据计算无论在实际应用还是科学研究中，都还在初期探索阶段，基于大数据得到的结论有的已经可以直接指导人们的社会实践，但也有不少还只是起到一定的参考作用，并不能取代传统的方法。

[7] ALPAC 是美国国会为调查 MT 而成立的“自动语言处理咨询委员会”的英文名首字母缩写。

析为词和短语结构，然后再按照目标语言的语序和结构要求，重新拼装，生成译文。但是，当这些针对小规模语言实例“表现良好”的理论和方法遇到大规模真实语料时，无论是对原文的分析，还是对译文的生成，人总结的理想的规则却远远无法胜任，人们对 NLU 的憧憬再次败倒在自然语言的无比复杂性面前。NLU 的大旗也逐渐易帜为 NLP（Natural Language Processing，自然语言处理），比起“理解”自然语言这样的目标，“处理”自然语言的信息，要务实得多。历史年轮很快转到了上世纪九十年代，伴随着互联网的迅速普及，主要以自然语言作为载体的海量数字化信息开始进入人们的生活。在这样的社会背景下，得益于计算技术的进步和大规模语言数据的易于获得，以统计方法为主导的 NLP 应用研究开始逐渐成为计算语言学学术会议和期刊论文的主角。从 1990 年 IBM 公司的 Brown 等人提出基于信源信道模型的统计机器翻译模型（Brown,1990,1993）到 2002 年 Och 提出基于最大熵的统计机器翻译方法（Och, 2002），在时隔半个世纪后，统计机器翻译再一次绕开了对语言结构的“理解”，让自然语言的翻译任务又一次回归到字符串信号处理（刘群，2003，宗成庆，2008）。2004 年 Och 加入 Google，基于统计的机器翻译借力 Google 的大规模双语对齐语料和并行计算平台，通过互联网开始为社会提供切实的翻译服务[8]。尽管跟以往基于规则的方法相比，翻译质量很难说有本质性的改观，但其开发周期短，维护成本低，支持语言多等诸多工程上的优势仍然广为业界称道。相比之下，传统的“先理解，再翻译”的所谓理性主义语言观不再是理所当然的信条。统计机器翻译的后来居上，让人们见识了计算机如何在大数据的平台上做到“不懂也能装懂”。为了近距离感受一下统计机器翻译方法的效果，下面不妨利用网上的三个在线机器翻译系统[9]，来做一个汉英翻译的小测试。

表 1：汉—英机器翻译示例

原文	伊拉克政府声明称，伊拉克政府坚持要求美国军队按照美伊驻军地位协议，在今年 6 月 30 日前从伊拉克城镇全部撤出，这一期限“不可延长”。
MT1	The Iraqi government stated said that Iraqi government insisted requested the American Army according to the US-Iraqi garrison status agreement, withdrew before June 30 from the Iraqi cities completely, this deadline “cannot postpone”.
MT2	He|she's station troops position agreement the Iraq government is declared saying the Iraq government persists in demanding USA troops according to US, comply with Iraq city and town before this June 30 all withdraw, this one time limit "is not allowed to prolong ".
MT3	Iraqi government statement said the Iraqi government insisted the United States military forces under the US-Iraq Status of Forces Agreement, in this June 30 to withdraw from all Iraqi cities and towns, the term "can not be extended."

对比表 1 中的三个机器翻译结果，不难发现，基于统计方法的 MT3 表现要更胜一筹。以原文中的几个语言难点：连续动词结构“声明称”“坚持要求”，专名“美伊驻军地位协议”，以及引语句“不可延长”的翻译结果来看，MT3 的译文结构都更准确，自然度也更高。基于规则的 MT1 系统的结果中出现了“stated said”“insisted requested”这样明显的语法错误。对“美伊驻军地位协议”这个专名的翻译，MT1 勉强可以接受，而 MT2 则完全没有翻译出来，而且还把其中的“伊”当成了第三人称代词，同时又无法确定其性别，因而译文中出现了“He|she”带上所有格标记“’s”的奇怪形式。这是基于规则方法的机器翻译系统更容易出现的问题。

[8] 目前 Google 在线翻译可以支持 66 种语言之间的互译。

[9] MT1 是国外的规则机器翻译系统；MT2 是国内的规则机器翻译系统；MT3 是国外的统计机器翻译系统。

尽管上面给出的基于不同方法的机器译文都算不上高质量，但总体来说，基于统计方法开发的机器翻译系统后来居上，超越现有的基于规则方法的机器翻译系统，已是不争的事实。计算语言学中发生这种研究范式的转变，并非偶然，而是有其深刻原因的：

（1）社会已经全面进入互联网时代。这个时代的特点是信息量大，信息传播速度快。自然语言的活跃程度远远高于以往任何一个时期。这就意味着语言字符本身的不确定性在增强[10]。这种情况对基于理性主义的规则方法，是一个比较严重的挑战。而用统计方法来发现不确定性对象背后的概率性的规律，则更为适应互联网时代的这种特点。

（2）互联网规模的惊人增速为统计模型准备了海量的数据，为统计方法大展拳脚提供了充足的弹药。比如基于手工构建的 Wiki 百科文章和整个互联网的网页文献 ，研究人员已经获得了巨型知识库如 DBpedia，Freebase, Probase，WikiTaxonomy, YAGO 等，并且仍在继续扩大规模。以 Freebase 为例，库中目前包含了 39,732,785 个主题和 1,814,525,012 个事实。基于如此庞大的知识库，新型的计算机问答系统（QA）就有能力回答诸如“Which composer from the eternal city wrote the score for the Ecstasy scene?（哪位来自永恒之城的作曲家是《沉醉》一剧的作曲者？）”这样的刁钻问题。（Weikum et.al.,2012，Ferrucci et.al., 2010）

（3）计算机的能力主要表现在“记忆”和“搜索”，而不是创新和演绎推理。统计方法在机器翻译以及中文分词等技术上的成绩，可以理解为计算机依靠其强大记忆能力，在海量数据和恰当的统计模型两驾马车的辅佐下取得的成功。完全人工的规则在语言知识的概括度和层级的系统性等方面可以表现出简洁的美感，但在工程应用层面，却缺乏对真实语料的有效覆盖，缺乏对具体而微的词语共现信息的准确刻画。人工规则更多的是在“类”（type）的层面描述语言对象的性质，而基于大数据的统计方法则基本上可以接近甚至做到在“例”（token）的层面描述语言对象的分布、搭配、对齐等方面的性质。

在上述这些因素的综合作用下，随着近十年来机器学习（Machine Learning）热潮在 NLP 领域的推波助澜，自然语言作为计算机的信息处理对象，其自身的特殊性越来越被工程技术人员淡化，研究人员更多的是从工程效果，而不是从内在理据的角度去看待他们开发的 NLP 系统。一种观念似乎已成为工程师们的共识：即便是最时髦的语言学理论，在 NLP 中也起不到多少锦上添花的作用。但是，话又说回来，这种状况显然并不是 NLU 的理想主义者所愿意看到的。当工程师们津津乐道于 NLP 凭借统计模型、机器学习技术所取得的最新成就的时候，也不乏传统的计算语言学的拥趸开始反思这个学科的未来之路。如果只是在工程上而不是在科学研究上具有独立性，计算语言学岂不成了应用统计学的一个分支（Wintner, 2009）？要实现人工智能的终极理想 NLU，仅靠 NLP 工程上的进步显然是不够的，没有了科学根基的工程技术，其命运大概只能是“行之不远”。那么，计算语言学以及更基础的语言学研究前进的方向又在哪里呢？

三 汉语研究的未来之路

本文并不想冒险去预测未来，但从过去的问题出发去探索未来之路总不是坏事。反观过去半个世纪计算语言学的发展历程，其实不难看到关于语言的理论研究的问题所在：（1）理论语言学的关注点过于注重所谓的抽象的“语言能力”，而在一定程度上忽视了具体的“语

[10] 自然语言的不确定性体现在两个方面：一是原本就有不少语言单位有不确定性；二是近年来由网络而逐渐扩散到普通社会生活用语中的新兴语言现象有明显加快的趋势。前者的例子如：（1）斯诺登给**北京和华盛顿**出了外交难题——美国“家务事”考验中国。（2）**北京和华盛顿**的时差是 13 个小时。其中“北京和华盛顿”在例 1 中指中美两国政府，例 2 中指地理上的两个城市。这种不确定性在网络时代变得更为常见。后者的例子比如“**被毕业，被自杀，被就业，被代表，被失踪，被小康，被增长，被繁荣，被开心、被捐款、被健康**、……”等许多不合一般语法的“被××”构造，“百度百科”中甚至有一个条目叫“**被时代**”。这类新的语言现象涉及到语言中的字、词、句、篇各个层次。

言使用”。（2）过去的语言学建模中大都只看自然语言的终端语符序列，即语言成品，基本忽略了作为交际主体的人的能动性，以及在交际过程中除语言符号本身之外的其他非语言本体知识的作用。

针对上述第一个问题，可以说大数据时代的语言工程正是一个改进的方向。现在比以往任何时候都能更容易地获得丰富的语言资源。借助集群计算机强大的计算能力和选择适当的统计模型，就有可能从海量语言数据中挖掘出更符合语言真实使用情况的规律知识，这不仅可以促进语言学理论研究，也有助于语言研究成果更好地转化为信息处理产品。

针对上述第二个问题，未来的语言学研究应该更注重跟心理学、神经科学、脑科学、认知科学研究的互动，把注意力从仅仅盯在终端语符序列，拓展到也深入考察语言交际的心理过程，研究人类在概念组织、意义推理等能力上的内在认知机制。事实上，计算语言学领域近年来的热点研究方向“隐喻理解”“情感分析”等，也已经从应用需求角度把这些值得深入探索的问题摆在了研究者的面前。已经有学者注意到，从心理学角度对文本（语言）特征及其创作者（或说话人）所做的分析，可以在面向应用的计算模型中发挥积极作用。比如基于英语的一些心理学研究发现，心情沮丧的学生更多地使用第一人称；说话人更多使用抽象的表达方式（形容词比动词更抽象）描述他人行为特征时，可能意味着描述中带有更多偏见；人们在指称表达式中给出的信息往往比所需要的更多[11]，等等（Krahmer, 2010）。

上述这两个方面中，第一个方面可能更具体一些，因为这是大数据时代对语言学提出的直接的要求，同时这也是语言学工作者的份内之事。这个方面做好了，再去跟其他学科交叉结合，可能也会更容易一些，而且进行大规模语言工程建设的过程，同时也就是检验既有语言学理论的过程，在这个过程中，很可能也会提出新的理论问题。下面主要就这个方面简略谈两点看法。

第一，汉语的电子化的大规模语言资源的数量、类型多样性、易获得性等方面都还有待提高。跟英语的情况相比，这方面汉语目前仍有较大差距。以美国宾州大学的语言数据联盟（LDC）[12]为例，LDC 是英语语言资源（同时也包括很多其他语种）的大超市，不同的研究单位按照 LDC 的格式规范将自己语言资源提交给 LDC，由 LDC 统一发布、管理、销售（既有免费资源，也有收费资源）。从 1993 年成立至今，LDC 的语言资源规模已经达到 565 种（其中中文资源有 50 种），包括语料库、知识库、音频资源、视频资源等多种形式。中国中文信息学会仿照 LDC 的做法，在 2003 年成立了 Chinese LDC（中文语言资源联盟）[13]，目前语言资源规模仅 95 种。差距可见一斑。此外，随着语言类型学的研究不断深入，积累的语言数据不断增加，国外也出现了可以方便查询的世界语言在线数据库，其中 WALS（世界语言结构地图）[14]是一个典型代表，WALS 目前包含了 2678 种语言的 76,492 个数据点。有些常见特征在很多语言中都有对应的数据采集，比如关于“宾语和动词的语序”特征，就有 1519 种语言的数据包含在 WALS 数据库中。国内汉语方言研究和少数民族语言研究多年来也积累了很多纸面的和若干电子化的材料，但把这些材料大规模数据化，并且放在互联网上供学术界使用，还未曾见到。我国学者向来有治学首先应注重材料的传统。在大数据时代，语言材料的规模已远超昔日，要继承乾嘉学派以来的朴学之风，就应该群策群力，联合起来，尽快将汉语语言资源电子化，并加以系统整理，放到互联网上供学界和社会使用。

第二，大数据时代的汉语语言资源建设不仅追求“量”，同时也重视“质”。语言资源的“质”可以从多个方面体现，包括（1）语言范畴形式化；（2）语言数据专项化；（3）语言

---

[11] 这跟 Grice 的“信息足量”语用原则并不完全一致。

[12] http://www.ldc.upenn.edu/

[13] http://www.chineseldc.org/

[14] http://wals.info/languoid/lect/wals_code_mnd (世界语言在线地图网站关于汉语普通话的数据）

知识可视化。总的目标就是让大型语言数据库规范、好用。

语言范畴形式化是构建大规模语言资源的理论基础和工程基础，即提出一套元语言符号系统，严密地表达一个语言模型，从而可以内部一致地对语言对象（事实）进行标识。比如汉语的词类体系，短语结构分类体系，语义分类体系等，都可以加以形式化，并用相应的范畴标记来标注汉语的语料。以加工汉语树库（Treebank）为例，我们拟定了 17 个短语范畴标记和 95 个词范畴标记[15]，对 100 多万字的汉语真实语料进行了分词、词性标注、句法结构标注。在标注过程中，发现了一些用传统的短语结构语法理论难以描述的语言现象（比如“他这是想家想的”，其句法结构树就很难用现有的短语结构进行标注），这就促使我们重新思考原来的汉语句法理论设计。而在标注完成后，我们可以定量分析树库中各词类、短语类的分布情况，以及词类序列构成歧义结构的情况等，这些定量分析反过来也可以评价初始的词类划分理论框架是否合理，为汉语的理论研究提供参考（詹卫东 2012a,b，2013）。

语言数据专项化是语言资源工程建设不断深化和扩展的自然结果。为获得优质语言数据，人们已经开始建设各种不同性质适应不同需求的大规模专项语言数据库，比如中文输入法中应用的超大规模的领域词典，文本情感分析中应用的情感词典，面向对外汉语教学的汉语述补结构用法词典等等，都是语言资源中的专项数据库。跟早期的通用型语言数据库相比，这些专项数据库通常选择特定的语言对象，有相对单一的应用目的，因而有可能在资源规模、质量和易用性等方面达到更高的水平。

语言知识的可视化，目标是以形象生动的方式展现枯燥的数据及数据间的关联。无论是宏观层面还是微观层面的语言事实，如果可以通过可视化界面来呈现相应的语言事实，用户就更容易直观地把握。下面是我们正在构建的汉语述补结构数据库的两个可视化页面。

图 1：述语“吃”所带结果补语

图 2：计算机自动提取的“干净”的相关事件角色

[15] 标记集参见：http://ccl.pku.edu.cn/doubtfire/Projects/Treebank_Tags.pdf

图 1 中“吃”所带的补语词“饱、完、掉、好……”等是人工搜集的。按照其在大规模语料中出现的频次高低，安排它们离“吃”的位置远近。频次高的距离“吃”近，反之则远。通过这种“距离像似性”，可以体会“吃”搭配不同补语的能力差异。点击其中的补语节点“干净”可以弹出一个文本框，显示“吃-干净”这个述补结构的一些基本信息。进一步点击框中的“事件角色”，则可以显示“吃”“干净”各自的事件参与角色和二者共享的事件角色。图 2 中的词语就是“干净”的事件参与角色，这些词语是从 CCL 现代汉语语料库[16]（3.3 亿字）中用程序自动抽取的，凡是跟“干净”在同一个句子中共现的名词，都被抽取出来，按照其共现频次高低安排在图中的位置、词的颜色及字号大小。频次越高的词位置越靠中间、颜色越亮，字号越大。尽管自动抽取的结果中有不少误判，但因为数据量大，那些典型的跟“干净”共现频率高的名词（如“衣服、人、水、房间、……”）还是凸显出来了。

在大数据时代，语言学家担当着语言数据（知识）的挖掘者，整理者，呈现者的角色。作为一个汉语研究者，有责任去挖掘和发现新的、有价值的汉语事实，并作出尽可能详尽的描写和尽可能合理的解释。而且汉语语言学研究应更加开放，更加重视多学科的交叉和融合。这要求我们自觉的用更加多元的视角去看语言对象，像盲人摸象一样，从单个视角，我们可能只能了解对象的一个侧面，如果多一些视角，就可以提供关于研究对象的更为完整的画面，使我们有可能更接近真理一些。这种开放的研究态度，并非大数据时代的新鲜事物，语言学理论研究中也早有先例。比如语言学家借鉴信息论的思想，提出把语言中的重音位置跟语言成分所负载信息量的大小关联起来的理论（端木三，2007），就是以跨学科视角开展研究的极佳例证。现在我们已经迈入到大数据时代，开展交叉和融合型的汉语语言学研究有更好的条件，理应更加普遍。

参考文献

Brown, F. Peter, Cocke, John, Della Pietra A. Stephen, Della Pietra, J. Vincent, Jelinek, Fredrick, Lafferty, D. John, Mercer, L. Robert, and Roossin, S. Paul, 1990, A Statistical Approach to Machine Translation, In *Computational Linguistics*, 1990, Vol.16, No.2.

Brown, F. Peter, Della Pietra A. Stephen, Della Pietra J. Vincent, Mercer, L. Robert, 1993, The

[16] http://ccl.pku.edu.cn:8080/ccl_corpus

Mathematics of Statistical Machine Translation: Parameter Estimation, In *Computational Linguistics* ,1993, Vol.19, No.2.

Etzioni, Oren, Tuchinda, Rattapoom, Knoblock, A. Craig , Yates, Alexander, 2003, To buy or not to buy: mining airfare data to minimize ticket purchase price, In *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data mining*.

Ferrucci,David, Brown, Eric, Chu-Carroll, Jennifer, Fan, James, Gondek, David, Kalyanpur, A. Aditya, Lally, Adam, Murdock, J. William, Nyberg, Eric, Prager, John, Schlaefer, Nico, Welty, Chri, 2010, Building Watson: An Overview of the DeepQA Project, *AI Magazine*, Vol. 31, No.3.

Ginsberg, Jeremy, Mohebbi, H. Matthew, Patel, S. Rajan, Brammer, Lynnette, Smolinski, S. Mark and Brilliant, Larry, 2009, Detecting influenza epidemics using search engine query data, *Nature,* Vol 457, 19 February 2009.

Krahmer, Emiel, 2010, What Computational Linguists Can Learn from Psychologists (and Vice Versa)? In *Computational Linguistics*, 2010, Vol. 36, No.2.

Leetaru, H. Kalev, 2012, A big data approach to the humanities, arts and social science, *Elsiever Research Trends, Special Issue,* Vol.30, Sept. of 2012.

Mayer-Schönberger, Viktor and Cukier, Kenneth, 2013, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Eamon Dolan/Houghton Mifflin Harcourt, 中文版：盛扬燕 周涛 译《大数据时代》，浙江人民出版社，2013。

Mayer-Schönberger, Viktor, 2011, *Delete: The Virtue of Forgetting in the Digital Age*, Princeton University Press , 中文版：袁杰 译《删除：大数据取舍之道》，浙江人民出版社，2013。

Och, Franz Josef, Ney, Hermann, 2002, Discriminative Training and Maximum Entropy Models for Statistical Machine Translation, *Proceedings of ACL 2002*.

Rajaraman, Anand, Ullman, D. Jeffrey, 2011, *Mining of Massive Datasets*，Cambridge University Press，中文版：王斌 译，《大数据：互联网大规模数据挖掘与分布式处理》，人民邮电出版社，2012。

Weikum, G., Hoffart, J., Nakashole, N., Spaniol, M., Suchanek, F. M. and Yosef, M. A., 2012, Big Data Methods for Computational Linguistics, *IEEE Data Engineering Bulletin Special Issue on Data Management beyond Database Systems*, Vol.35, No.3.

Wintner, Shuly, 2009, What Science Underlies Natural Language Engineering? In *Computational Linguistics*, 2009, Vol. 35, No.4.

端木三，2007，重音、信息和语言的分类，《语言科学》2007 年第 5 期。

冯志伟，2010，《自然语言处理的形式模型》，中国科学技术大学出版社.

刘群，2003，统计机器翻译综述，《中文信息学报》2003 年第 4 期。

詹卫东，2012a，基於大規模中文樹庫的漢語句法知識獲取研究，第四届汉学国际会议，台湾中研院语言学研究所，2012.6.20-22， 台北。

詹卫东，2012b，从语言工程的角度看“中心扩展条件”与“并列条件”，《语言科学》2012 年第 5 期

詹卫东，2013，计算机句法结构分析需要什么样的词类知识——兼评近年来汉语词类研究的新进展，《中国语文》2013 年第 2 期。

宗成庆，2008，《统计自然语言处理》，清华大学出版社。

致谢：本文研究工作得到教育部人文社会科学研究项目规划基金项目“现代汉语述补结构网络数据库的构建与应用”（项目编号：12YJA740104 ） 和国家社科基金项目“语言知识资源的可视化技术研究”（项目编号：12BYY061）的资助，谨致谢忱。

大数据时代的汉语语言学研究

摘要 借助互联网的迅猛发展，当今社会已经进入“大数据”时代。文章通过回顾计算机科学与语言学的交叉学科——计算语言学的发展历程，从一个侧面揭示了大数据处理对科学研究的冲击和影响，并在基础上探讨汉语语言学研究的未来之路。文章认为，在当前的时代背景下，应该更加注重语言工程的研究和开发：汉语大规模语言资源的数量、类型、易获得性均有待提高；汉语语言资源建设应努力实现语言范畴形式化、语言数据专项化和语言知识可视化。汉语语言学的研究应更加开放、更具多元化视角、更加注重多学科的交叉和融合。

关键词 大数据 计算语言学 汉语语言学 语言资源 形式化 可视化

Chinese Linguistics In The Era of Big Data

Abstract: The central challenge of our times is that data is growing at extraordinary rates because of the ubiquitous Internet. In the so-called "big data era", the research paradigm and methodology of many disciplines are changing rapidly. This paper illustrates what has already happened in an interdisciplinary research area, Computational Linguistics under the influence of big data. By learning from the experiences of its neighbors, this paper puts forward a proposal that researchers of Chinese Linguistics should pay more attention on language engineering to enhance the development of large-scale electronic Chinese linguistic resources, including the amount as well as type diversity and accessibility. In order to achieve that goal, Chinese linguists should make endeavors to realize the systematical formalization of Chinese linguistic notions, to build more linguistic databases for more specific purpose and to render visualization of large-scale linguistic data for rapid, easy and high effective use. A multidisciplinary perspective should be encouraged actively for Chinese linguists' adapting to the new environment in the era of big data.

Keywords: big data, computational linguistics, Chinese linguistics, language resource, formalization, visualization