LINGUISTICS 语言学博士文库

系统功能语言学在自然语言处理中的知识表示研究

Knowledge Representation of Systemic Functional Linguistics in Natural Language Processing

李学宁 李向明 宋孟洪◎著

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

LIGUISTICS|语言学博士文库
广西民族大学科研基金资助课题（引进人才科研启动项目2017MDSKRC067）

系统功能语言学在自然语言处理中的知识表示研究

Knowledge Representation of Systemic Functional Linguistics in Natural Language Processing

李学宁　李向明　宋孟洪◎著

上海交通大学出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

内容提要

随着计算机的诞生，自然语言处理逐步成为了语言学的一个重要应用领域。由于语言学知识一般用自然语言进行表述，有必要采用一些形式化方法将其重新表示出来，才能为计算机识别和处理。因此，知识表示就成了语言学和计算机科学中的一个共同前沿课题。在此背景下，本书对系统功能语言学的知识表示进行跨学科的研究。主要内容包括：① 按时间顺序，依次考查系统功能语言学在机器翻译（含机器词典）、自然语言生成、多模态语料库中的应用情况；② 在此基础上，归纳总结系统功能语言学在应用过程中所采用的知识表示方法；③ 针对一种最常用的知识表示方法——系统网络，考查它在表示能力和可实现性方面存在的缺陷，并基于系统论和计算机技术的最新发展，提出一个复杂系统网络型式。本书可供语言学专业博士生、硕士生以及对此领域感兴趣的学者阅读参考。

图书在版编目（CIP）数据

系统功能语言学在自然语言处理中的知识表示研究/李学宁，李向明，宋孟洪著. —上海：上海交通大学出版社，2018

ISBN 978-7-313-20120-1

Ⅰ. ①系… Ⅱ. ①李… ②李… ③宋… Ⅲ. ①自然语言处理—研究 Ⅳ. ①TP391

中国版本图书馆 CIP 数据核字（2018）第 203821 号

系统功能语言学在自然语言处理中的知识表示研究

著　　者：李学宁　李向明　宋孟洪
出版发行：上海交通大学出版社
邮政编码：200030
出 版 人：谈　毅
印　　制：上海天地海设计印刷有限公司
开　　本：710 mm×1000 mm　1/16
字　　数：106 千字
版　　次：2018 年 10 月第 1 版
书　　号：ISBN 978-7-313-20120-1/TP
定　　价：48.00 元

地　　址：上海市番禺路 951 号
电　　话：021-64071208
经　　销：全国新华书店
印　　张：8
印　　次：2018 年 10 月第 1 次印刷

序

今悉李学宁教授的《系统功能语言学在自然语言处理中的知识表示研究》一书即将付梓，十分高兴。我和李学宁教授早在 2001 年就认识了。当时我们正在计划写作《功能语言学与外语教学》一书，其中一个部分需要找人合作完成，那就是如何把系统功能语言学用于测试。当时，既懂系统功能语言学又做测试的人不多，李学宁教授正好写了一篇这样的论文让我看，我想这正是我要找的人，就和他讨论合作一事，他欣然接受。后来，他到上海交通大学攻读计算机专业的博士学位，并有意继续深造，做博士后研究。我当时可以在山东大学带博士后，所以，他就毅然辞去了系主任的职务，转而做博士后研究。

在这种情况下，他就具备了跨学科的优势，既懂系统功能语言学，又是计算机专业的博士生，所以，就和他协商是否他还愿意继续做计算机研究，探讨如何把系统功能语言学理论应用于计算机科学，发展系统功能计算语言学。他说他的想法和我的是一致的，所以，就开始做这方面的研究，经过两年多的辛勤刻苦的研究，他按规定完成了各项学习和研究任务，顺利完成了博士后研究出站。本书是该博士后研究项目的主要成果。

作为一门适用语言学，系统功能语言学自诞生之日起就开始探索如何应用于计算语言学，并在应用的过程中不断发展自己的理论。系统功能语言学开始于阶与范畴语法的研究（Halliday，1961）。Halliday 在此之前就应用“阶”与“范畴”的基本观点对机器翻译和机器词典的语言学原理进行了初步的探索。并且，他针对机器翻译中对等翻译缺失的情况提出了“同义词词库法”（Halliday 和 Webster，2007：6－41）。在 20 世纪 60 年代，Henrici 基于

Halliday 的系统思想开发了一个计算机程序，它能够对一个系统网络中的不同选项进行自动选择（Matthiessen & Bateman，1991：18）。1972 年，Winograd 研制了一个自然语言处理系统——SHRDULE。在这个系统中，他首次采用了 Halliday 的系统语法来建立一个比较全面的英语语法。在其后 20 年中，一系列基于系统功能语言学的语篇生成系统应运而生，比较著名的有 PROTEUS、PENMAN、SLANG、COMMUNAL 等。迄今为止，系统功能语言学已经成为语篇生成系统中应用最为广泛的语言学理论之一。进入 20 世纪 90 年代以后，系统功能语言学开始积极地应用于语料库语言学的研究以及各类语料库的建构，其最终目的仍然是服务于自然语言处理的需要。与此同时，Halliday（Halliday & Webster，2007：157－189）还开展了对口语语料库的研究。Bateman（2008）、Baldry（2010）等进一步研究了多模态语料库的建构，并成为系统功能语言学在当代自然语言处理应用研究中的一个重要动态与发展趋势。针对人工神经网络研究与遗传算法、模糊逻辑结合起来发展智能计算（Intelligent Computing）的趋势，Halliday 从系统功能语言学的角度探讨了语言的模糊性，并初步论述了系统功能语言学在智能计算中的应用前景（Halliday & Webster，2007：193－267；杨才英，2007）。

总之，系统功能语言学在人类计算机发展的每个关键时期都做出了积极的贡献，主要应用于机器翻译和机器词典、自然语言生成和（多模态）语料库；而这些应用领域与计算语言学本身的发展趋势基本一致。由此可以说：计算语言学是系统功能语言学的一个重要应用领域；而在计算语言学中的应用反过来深刻地影响了系统功能语言学的发展轨迹，并推动了其自身的理论建设，尤其是相关的形式化研究的开展。

本书聚焦于如何用系统功能语言学理论在自然语言处理中进行知识表示，即采用一些形式化方法将其重新表示出来，为计算机识别和处理提供工具。本书按照跨学科的思路，采用历史文献法和形式化方法。通过文献法，全面梳理系统功能语言学在一些自然语言处理主要领域的应用情况。而形式化方法就是知识表示方法，它是系统功能语言学与计算语言学发生关联的“玄关”。通过研究发现：系统功能语言学在本质上是一种面向自然语言处理且在该领域中具有广泛应用价值的语言学理论。在应用过程中，系统

功能语言学者提出并发展了系统网络，从而丰富了计算语言学和人工智能中的知识表示方法。

本书在国内首次全面、系统地研究了系统功能语言学在自然语言处理中的应用情况及相关的知识表示方法，使大家初步认识系统功能语言学的计算语言学的理论、方法和特点。它能够在一定程度上纠正人们对于系统功能语言学的一种误读，即该语法不重视甚至反对形式化研究；也有助于从一个更加全面和客观的角度来审视系统功能语言学的发展轨迹——它的发展基本与计算语言学的发展同步，可以使大家更好地理解 Halliday 等人在众多的语言学理论观点中进行选择的技术动因。希望本书得到广大读者的喜爱。

张德禄
同济大学教授、博士生导师
2018 年 3 月 7 日
于上海文化花园香榭丽苑

前　言

随着计算机的诞生，“自然语言处理”逐步成为语言学的一个重要应用领域。由于语言学知识一般用自然语言进行表述，有必要采用形式化方法将其重新表示出来才能为计算机识别和处理。因此，“知识表示”就成为语言学和计算机科学中的一个共同前沿课题。

在此背景下，本书对系统功能语言学的知识表示进行跨学科的研究。主要内容包括：① 按时间顺序，依次考查系统功能语言学在机器翻译（含机器词典）、自然语言生成、多模态语料库中的应用情况；② 在此基础上，归纳总结系统功能语言学在应用过程中所采用的知识表示方法；③ 针对系统功能语言学最常用的知识表示方法——系统网络，考查它在表示能力和可实现性方面存在的缺陷，并基于系统论和计算机技术的最新发展，提出其发展趋势为复杂系统网络型式。

在研究过程中，本书采用的主要方法有历史文献法和形式化方法。通过文献法，全面梳理系统功能语言学在一些自然语言处理主要领域的应用情况。而形式化方法就是知识表示方法，它是系统功能语言学与计算语言学发生关联的“玄关”。

通过上述研究，所获得的一个重要发现是：系统功能语言学在本质上是一种面向自然语言处理且在该领域中具有广泛应用价值的语言学理论。在应用的过程中，系统功能语言学者提出并发展了系统网络，从而丰富了计算语言学和人工智能中的知识表示方法。

在国内，这项研究较系统地研究了系统功能语言学在自然语言处理中的应用情况及相关的知识表示方法。它能够在一定程度上纠正人们对于系统功能语言学的一种误读，即该派理论不重视甚至反对形式化研究。此外，也有助于从一个更加全面和客观的角度来审视系统功能语言学的发展轨迹：它的发展与计算语言学基本同步。正因如此，我们才能更好地理解 Halliday 等人在众多的语言学理论观点中进行选择的技术动因。

目　录

第1章 绪 论

1.1 研究背景与意义

1.1.1 研究背景

近现代西方语言学史上有两个重要的里程碑。

第一个是《普通语言学教程》的出版，标志着语言研究在性质上已经完全实现了从传统语法、历史语言学到现代语言学的嬗变。在该书中，Ferdinand de Saussure 提出了一系列创新性的概念、分析方法和基本原则，对 20 世纪的布拉格学派、哥本哈根学派、美国结构主义学派、转换生成语法和伦敦学派等均产生了深远的影响（Matthiessen 和 Bateman，1991：77－78；张同俊，2010）。

另一个分水岭是计算机的诞生，它标志着语言的研究已经正式进入了一个以定量和可计算性为主要特征的电子时代（戴炜栋、张爱玲，1999）：

（1）语言学和数学之间具有紧密的关系。Saussure 本人就指出语言之间的数量关系可以用数学公式表示出来；而语言系统就好比是一个几何系统，它可以归结为一些有待证明的定理。1933 年，美国语言学家 Leonard Bloomfield 就提出数学不过是语言所能达到的最高境界。在 1945 年之前，

许多数学家如英国的 A. de Morgan、美国的 T. C. Mendenhall、加拿大的 E. Varder Beke、中国的陈鹤琴，尤其是俄国的 Markov，Andrei Andreevich 等用数学方法对语言进行了实际的研究，并推出了一批重要的成果。然而，这些工作对于当时的语言学研究并没有产生显著的影响。绝大部分语言工作者仍孤立于数学之外，沿着自己的传统道路迟缓地发展着(冯志伟，2011a：1－2)。

(2) 早在计算机诞生之前，英国数学家 A. M. Turing 就天才地预言了计算机与自然语言之间的紧密联系："我们可以期待，总有一天机器会同人在一切的智能领域里竞争起来。但是，以哪一点作为竞争的出发点呢？这是一个很难决定的问题。许多人以为可以把下棋之类的极为抽象的活动作为最好的出发点，不过，我更倾向于支持另一种主张，这种主张认为，最好的出发点是制造出一种具有智能的、可用钱买到的机器，然后，教这种机器理解英语并且说英语。这个过程可以仿效小孩子说话的那种方法来进行(冯志伟，2010：10)。"

(3) 自计算机诞生以后，在一大批语言学家、数学家、计算机专家等不同领域学者的共同努力下，产生了两门新兴的前沿交叉学科——计算语言学(Computational Linguistics)和人工智能(Artificial Intelligence，AI)。计算语言学通过建立形式化的数学模型来分析、处理自然语言，并在计算机上用程序来实现分析和处理的过程，从而达到以计算机来模拟人的全部或者部分语言能力的目的(俞士汶，2003：2)。而人工智能则试图了解智能的实质，生产出一种新的能以人类智能相似的方式做出反应的智能机器。其主要研究领域有机器人、语音识别、图像识别、自然语言处理和专家系统等。

显然，这两门学科都具有一个共同的研究方向和领域——自然语言处理(Natural Language Processing，NLP)，[①]它的研究对象是自然语言，即人们在日常工作和生活中所使用的语言，其最终目的是为了实现人与计算机之间的自然语言通信，因此归属于计算机科学。经过几十年的发展，它形成

① 也有部分专家认为自然语言处理已经上升为一门相对独立的学科，它主要用于说明方法，侧重于工程(江铭虎，2006：1)。

了自然语言理解(Natural Language Understanding)和自然语言生成(Natural Language Generation)两大分支。前者指的是计算机理解自然语言文本的意义,后者指计算机以自然语言文本来表达给定的意图和思想。两者共同构成了人、机自然语言通信的一个完整过程。因此,我们在涉及自然语言处理的研究中,往往对计算语言学和人工智能两门学科不加以严格的区分。在更多的时候,我们将使用"计算语言学"这一术语。

在此背景下,语言学理论,尤其是试图应用于自然语言处理的语言学理论面临一个共同的问题:语言学知识一般是通过自然语言形式,即人们日常所使用的语言进行阐述;而这不能为计算机所直接识别和处理。因此,我们有必要采用一定的形式将这些知识表示出来。这样一来,"知识表示"(Knowledge Representation)就成为语言学和计算机科学共同关注的前沿课题。

1.1.2 研究意义

本书关注的是系统功能语言学在自然语言处理中的应用及其知识表示方法。其意义在于:

(1) 系统功能语言学是一门"适用语言学"(Applicable Linguistics),具有广泛的应用领域(胡壮麟,2007)。在各种应用领域中,M. A. K. Halliday尤为重视自然语言处理,并明确提出能否将语言学理论直接应用于计算语言学或人工智能是检验语言理论是否正确、是否完善的重要手段,同时也是使语言理论发挥更大作用的大好机会(朱永生、严世清,2001: 12)。

为了实现上述目标,就有必要研究系统功能语言学与计算语言学之间的接口——知识表示方法,即将原本用自然语言描述的语言知识转换为一种可以为计算机识读的方式重新表示出来。从某个角度来说,能否直接采用自然语言处理中行之有效的知识表示方式,就成为检验系统功能语言学能否直接应用于计算语言学的一个重要评判标准。

(2) 系统功能语言学的一个重要应用领域是自然语言生成。时至今

日，它已经成为语篇生成系统中应用最为广泛的语言学理论（邵军力、张景、魏长华，2003：251）。

然而，我们对于系统功能语言学在这些语篇生成系统中的应用情况还缺乏深入的了解，尤其是还茫然于系统功能语言学在这些系统中具体的知识表示方式。通过本研究的开展，我们可以总结和归纳这方面的研究进展，从而不断改进并最终找到一个更适合于系统功能语言学的知识表示方式。

更为重要的是，我们有必要进一步了解系统功能语言学在计算语言学中其他领域的应用情况，例如机器翻译、（多模态）语料库等，从而对于两个学科之间的关联及其发展趋势有一个更加全面的认识。

（3）从20世纪80年代开始，国外语言学理论的研究已经出现了一种新的动态，即各种应运而生的语法体系往往是直接服务于自然语言处理的需要，例如Gerald Gazder的广义短语结构语法（王宗炎，1985）、Martin Kay的功能合一语法（冯志伟，1991）、Joan Bresnan和Ronald Kaplan的词汇功能语法等（俞如珍、金顺德，1994：438－474）。

这种局面给我们带来了诸多启迪和反思：语言学知识的创新是继续沿用传统的自然语言方式来进行描述，还是从一开始就尽可能地采用一种适合于自然语言处理用途的知识表示方式？显然，我们这项工作有助于推动传统的系统功能语言学研究向计算系统功能语言学（Computational Systemic Functional Linguistics，CSFL）的发展。①

总而言之，这项工作的开展有助于人们全面地了解系统功能语言学在自然语言处理中的应用情况。在此基础上，反过来对系统功能语言学理论本身的科学性进行一个历时的、全面的评估，并从一个新的角度来探讨系统功能语言学与其他语言学理论，尤其是计算语言学之间相互借鉴的发展趋势。

---

① 计算系统功能语言学指系统功能语言学在计算机上的知识表示，它是系统功能语言学与计算语言学之间的交叉分支学科。“第一届世界计算系统功能语言学国际大会”于2005年7月在悉尼大学举行。

1.2 国内外研究现状述评

1.2.1 国外研究现状

在这个部分，我们有必要先勾勒出计算语言学的发展概貌，然后进一步梳理系统功能语言学在自然语言处理中的应用历程。

1.2.1.1 计算语言学的发展概貌

计算语言学的发展大致分为三个主要的时期(冯志伟，2011b)：

(1) 萌芽期(从20世纪40年代至50年代末)。在萌芽期，一般认为最重要的四项理论成果分别为Markov的马尔可夫模型、Turing的算法计算模型、Claude Elwood Shannon的概率和信息论模型、Avram Noam Chomsky的形式语言理论。他们的工作为计算语言学的诞生与发展奠定了坚实的理论与技术基础。

在此期间，最引人注目的应用领域是机器翻译。1954年，IBM在全世界首次采用计算机进行了第一次机器翻译实验，紧接着苏联、英国和日本等国也进行了类似机器翻译实验，出现了一股机器翻译的热潮。然而，机器翻译不久就遇到了难以克服的语义障碍(Semantic Barrier)。随着ALPAC报告的发表，机器翻译在全世界范围内的热潮顿时消失了，呈现出了一片萧条的局面。

(2) 发展期(从20世纪60年代中期至80年代末)。在发展期的研究中出现了两种重要的趋势：一是重新评价了有限状态模型，并成功地将其应用于音系学、形态学和句法学的研究之中；二是重拾了经验主义的研究方法。这种方法的理论基础是信息论，它将语言事件赋予概率。研究者们认为人的知识通过感官输入，经过一些简单的联想与通用化的操作而获得。因此，他们认为计算语言学的研究对象是实际的语言数据，并从大量的语言数据中获得语言知识。显然，这种主张是对Chomsky理性主义的一种反叛，即