汉语实词语法功能考察及词类体系构建

HANYU SHICI YUFA GONGNENG KAOCHA JI CILEI TIXI GOUJIAN

徐艳华\著

汉语实词语法功能考察及词类体系构建

HANYU SHICI YUFA GONGNENG KAOCHA JI CILEI TIXI GOUJIAN

徐艳华\著

中国社会科学出版社

图书在版编目（CIP）数据

汉语实词语法功能考察及词类体系构建/徐艳华著．
北京：中国社会科学出版社，2007.11
ISBN 978-7-5004-6545-4

Ⅰ.汉… Ⅱ.徐… Ⅲ.①汉语-实词-语法-研究
②汉语-词类-研究 Ⅳ.H146.2

中国版本图书馆CIP数据核字（2007）第175399号

出版策划 任 明
特邀编辑 李晓丽
责任校对 李 莉
封面设计 典雅设计
技术编辑 李 建

出版发行 中国社会科学出版社
社 址 北京鼓楼西大街甲158号 邮 编 100720
电 话 010-84029450（邮购）
网 址 http://www.csspw.cn
经 销 新华书店
印 刷 北京奥隆印刷厂 装 订 一二零一印刷厂
版 次 2007年11月第1版 印 次 2007年11月第1次印刷
开 本 880×1230 1/32
印 张 5.375 插 页 2
字 数 152千字
定 价 23.00元

凡购买中国社会科学出版社图书，如有质量问题请与本社发行部联系调换

作者简介

徐艳华，女，1976年生，山东烟台人，2003年获烟台师范学院文学硕士学位，2006年获南京师范大学文学博士学位，现为鲁东大学汉语言文学院讲师。主要从事现代汉语语法和计算语言学研究。先后发表学术论文多篇，参编词典三部，参加科研项目多项。

内容简介

本书主要以计算机为服务对象，以语料库为信息源，采用统计并辅以内省的方法，详细考察了名词、动词、形容词和副词中高频的3514个词的语法功能，依据统计结果对这些词进行了分类，并从理论和实践两个方面对分类结果进行了验证。研究表明：从理论上来看，这种分类方法能够客观地反映现代汉语中词的多功能现象；在实践方面，该书的分类结果在外显式歧义结构的消解和基本名词短语的识别方面都起到了很好的作用。

出版策划：任 明

封面设计：典雅设计

中文摘要

随着信息社会对信息自动化处理要求的不断提高，越来越需要计算机能对自然语言进行深层分析，比如文本校对、机器翻译、自动文摘等等，这就要求我们为计算机提供尽可能多的语言知识尤其是语法知识，使其能够进行正确的分析进而做到真正的“理解”。现有的汉语语法体系，可以说已经研究得比较深入了，但其主要是面向人的，面对计算机这个新的交际伙伴，其研究成果还是显得有些粗糙。因此，要真正实现信息自动化处理，必须加大汉语语法研究的力度，加细语法刻画的颗粒度，构建真正适合计算机处理自然语言需要的汉语词类体系。本书正是在这样的服务宗旨下展开研究的。

本书研究主要从如下三个方面展开：

1. 以语料库为信息源，采用统计并辅以内省的方法，详细考察了名词、动词、形容词和副词中高频的3514个词的语法功能，构建了语法功能信息库，并以信息库中统计的数据为依据，彻底贯彻“按照词的语法功能划分词类的标准”，依据“句法功能完全相同即为一类”的原则，对3514个常用词进行了分类，最终分出676类，其中，一词一类的有364类，两词一类的有107类，两词以上同类的有205类，对于一词一类和两词一类的，我们放到词典里描述，剩下的205类就是新实词词类体系所包含的类别。

2. 从句法功能复杂度和分类结果两个方面对现有的词类体系和我们构建的词类体系进行了对比研究。研究结果表明，只有

10.1%的词在两种体系中的句法功能复杂度是相同的，而有84.1%的词在旧体系中的句法功能复杂度是高于新体系的。这说明现有的词类体系夸大了汉语中词的多功能现象，归属同一类别的词，不管是有1项功能还是具有10项功能都给以相同的标记显然有失偏颇，不能客观地反映汉语的实际情况。另外，从分类结果看，新体系中存在一个词类包含旧体系中多个词类的情况，这说明旧体系通过找“特点”来为词归类的做法不可取，必须在详细考察的基础上才能做到“词有定类”。

3. 从清华大学100万字的汉语句法树库中提取了11206个v + v序列和10081个v + n序列进行了考察。前一种序列匹配后只有一种句法关系的有2679种组合模式共10296个实例，正确率分别为82.9%和91.9%。后一种序列匹配后只有一种句法关系的有1462种组合模式共7189个实例，正确率分别为70.7%和71.3%。其他有两种以上句法关系的模板，尽管不能确定具体实例中究竟是哪种关系，但相对于旧体系来说，其歧义的数量大大减少了。统计结果表明，新体系在外显式歧义结构的消解和基本名词短语的识别方面都起到了很好的作用。

关键词： 句法功能；词类体系；句法分析；歧义消解

Abstract

The deep analysis of the natural language, such as grammar checking, machine translation and text summarization, is more and more required to fit the rapid development of computer's automatic information – processing. The computer should be provided with the language information especially that of the grammar so as to ensure the correct analysis and automatic understanding. Generally speaking, the Chinese grammar system available is human – oriented, that is to say, it is good enough for human beings though, it is rather rough for the computer. But the realization of the automatic information – processing results from the deepening of the Chinese grammar researches and the specification of the grammar characterization, especially the reconstruction of the Chinese word class system suitable for computer natural language processing. Taking this as the goal, this dissertation made the following discussions.

The first part, based on a corpus and taking statistic and introspective methods, makes a detailed survey of 3514 high frequency words' grammatical functions, which are nouns, verbs, adjectives and adverbs, and it establishes a database of grammatical function information. Besides, it classifies the 3514 high frequency words into 676 classes, in thorough accordance with the guideline of grammatical function. The result is that there are 364 classes with only one word, 107 classes with two words, and 205 classes with more than two words, among which, the

first two classes are attributed to the lexical base, and the 205 classes are formed as a new substantive word class system .

The second part compares the word class system made by other researchers and our new one in terms of the syntactic function complexity and classifying results. As for the function complexity, 10. 1% are the same while 84. 1% of the old one is more complicated than the new one. This shows that the old one exaggerates the multi – functional phenomena in Chinese word system. As for the classifying result, that one word class in the new system contains more than one word classes in the old system exists. This shows that the conclusion of one word with a particular word class can only be made on detailed survey and exploration instead of feature finding.

The third part observes 11, 206 V + V phrases and 10, 081 V + N phrases from the Chinese syntactic tree – diagram corpus made by Tsinghua University to testify the application of the new word class system. The result shows that the ratio of correctness of the first kind of phrases is 82. 9% and 91. 9%; 70. 7% and 71. 3% of the second kind. Compared with the old system, the new system can greatly dissolve the ambiguity caused by the same form with the different deep structures and can easily recognize the basic noun phrases.

Key words: syntactic function; word class system; syntactic analysis; ambiguity dissolving

目　录

前　言

一　课题的提出

本书的研究工作是以大规模语料为基础，在充分考察每一个实词语法功能的基础上自底向上地重构现代汉语实词词类体系。这是一种彻底按照词的句法功能标准重构汉语实词词类体系的尝试，同时也为计算机进行自动句法分析提供更详细完备的句法信息，以期减少句法分析中的结构歧义现象。

在过去的语法研究中，有关现代汉语词类问题一直是语言学界关注的焦点，诸多语言学家曾投入大量精力进行了深入研究，形成了现代汉语词类体系，但其研究的服务对象主要是面向人的。随着计算机科学技术飞速发展以及信息社会对信息自动化处理的要求不断提高，语法研究的应用对象由过去面向人发展到现在不仅面向人还面向计算机，而且后一个方面显得越来越迫切和重要。鉴于此，本书的研究工作由以往主要是面向人的语法研究转向主要面向计算机。目前的信息处理技术，比如文本校对、机器翻译、自动文摘等越来越多地需要对自然语言进行深层分析。开发这类应用系统，就要求我们为计算机提供尽可能多的有关自然语言知识和非语言知识，前者又包括句法知识、语义知识乃至语用知识等等。

衡量一个自然语言处理系统的水平，可以看它处理到语言单位中的哪个层级，同时更要看它对不同性质的语言知识掌握到什么程度。无论是比较传统的基于规则的处理策略，还是基于统计的方法，在对语言知识的需求这一点上实际都是一致的。所不同的是，

采用基于规则方法的研究者一般诉诸专家的理性知识，由人根据已有的知识储备来对语言知识进行抽象，比如根据一个词能作主语、宾语、定语、中心语等功能给以名词“n”的标记；而采用基于统计方法的研究者一般求助于计算机对大规模语料库进行统计分析，由计算机来抽象出语言知识，比如以一定的数据结构记录统计结果等。两种研究方法孰优孰劣，不能笼统判断，只能跟具体的应用目标结合起来，由实践结果来评价。统计方法已经在像语音识别、自动分词和词性标注这样相对浅层的自然语言处理中有不俗表现，但在深层分析方面，比如分析句子的树结构或者句法成分的语义关系等领域还没有显示出特别的优势。于是又有学者提倡把两种方法结合起来使用，比如通过统计，给出带有概率值的规则。在我们看来，无论采用哪种方法，首先都要求人自身先对自然语言有深入的了解。就规则方法来讲，这一点是显然的；就统计方法来讲，虽然不那么明显，但道理也是一样的。现有的对自然语言深层知识的统计，一般是建立在经过标注的熟语料库基础上的，而从生语料库到熟语料库，就具体的加工方式而言，当然有人工方式，也有计算机自动加工方式或者人机互助的方式等等，但加工什么内容，标注哪些信息，仍然取决于人对自然语言的认识。

具体到中文信息处理方面，如果从宏观上以处理对象的单位大小为指标来看，中文信息处理在汉语的字处理方面已经比较成熟，词处理阶段的形式方面比如说分词和词性标注等已经取得了一定的成果，并且基本上可以达到应用的目的，词的意义处理、词组和句子的结构处理还比较薄弱，至于篇章处理和各层次的环境处理还在摸索中。目前的研究主要在句子一级展开，包括自动分析句子的内部句法关系、给出结构成分间的语义关系等不同深度的分析。单就自动句法分析来看，作为汉语研究者需要考虑这样两个问题：一是从自动句法分析的需要考虑，重点应该为计算机提供哪些语言知识；二是现有的技术条件和语言学研究水平又能够为计算机提供多少。

基于对上述两个问题的思考，选择了本研究课题。

对于第一个问题的回答，主要是根据中文信息处理已有的研究成果和从目前的实际需要出发，初步确定了本课题研究的主要内容。一般来说，自动句法分析的操作对象是句子或短语的词类标记序列，“客观的句法分析只能根据词类的标记序列来推知句法结构，如果词类问题没有解决好，或者词类和句法分析脱钩，那就无法根据词类序列去分析句法结构，这样就会影响整个语法体系的科学性和实用价值。”① 由此看来，要进行自动句法分析，除了要为计算机提供必要的语义知识、词语搭配知识以及关于客观世界的知识外，更重要的是要为其提供比较完备的语法知识，尤其是语法基础的词类知识。从20世纪80年代中后期开始直到现在，研究人员已经在汉语词语的语法功能分类和属性特征描述方面开展了卓有成效的工作，希望为计算机分析汉语句子结构打下一个很好的基础。但实际上，这个“基础”并不能真正满足计算机进行自动句法分析的需要，因为其分类并不是在详细考察每个词的语法功能的基础上进行的，所以目前迫切需要解决的问题是，对汉语实词的句法功能进行全面系统的考察，在这个考察过程中得到的结果，不仅可以检验以往对词的语法知识的概括是否合适，从而进行相应的调整；而且可以根据统计分析的结果为自动句法分析构建比较合理的实词词类体系，这样的知识对于没有任何隐含知识的计算机来说是进行句法分析必不可少的。只有尽可能地把每类词的句法功能描述清楚，为计算机提供更加精细和完备的词类体系，才能指导它分析出正确的句子结构，给出正确的语义解释。而从发展趋势来看，越来越多的高级自然语言处理应用系统的研究与开发，诸如信息提取、机器翻译等，也都离不开这样的语法知识的支持。

对于第二个问题的回答，则主要是结合我们对目前现代汉语词类体系以及自动句法分析的具体需要这两方面的认识，大致确定了

① 胡明扬：《词类问题考察》，北京语言文化大学出版社1996年版，第1页。

本课题研究应该追求的合理目标。从历史上看，汉语的词类体系是以印欧语语法的词类体系为蓝本的。尽管经过几代语言学家的研究，根据汉语的实际情况作了一些局部调整，比如增加了量词、助词和语气词，从形容词中分出区别词等等，这些局部调整的确不乏闪光之处，但是基本的格局没有改变，依然无法摆脱模仿的痕迹。由于始终摆脱不掉印欧语词类体系的羁绊，所以最终导致汉语词类划分并不是严格地按照词的语法功能来进行的。特别是实词分类，似乎名词、动词、形容词是生来就有的、不必加以验证的词类。尽管说语言学界对词类问题进行过几次大讨论之后逐步达成共识，认识到语法功能是词类划分的唯一标准，但在实际操作中，这一标准并没有被真正彻底地贯彻。就现有的词类体系看，“不管哪种类型，也不管是哪个版本，无一不带有先验性。它们都不是对客观存在的词进行全面分析和全面归纳的产物，而是先由语法学家所构拟然后又由语法学家加以解说的框架，这样的框架必然带有语法学家的成见和缺陷。”① 具体说来有这样几个方面：（1）每个词类到底有哪些语法功能，这一点很不明确，一般的语法书上仅列出几条“语法特征”。例如，说名词可以受数量结构修饰；不能受副词修饰；可以作主语、宾语等等，但即便是这几条特征也往往缺乏普遍性。（2）属于同一词类的词，其语法功能可能有很大差异。例如，“领导”可以作主语、宾语、定语、体词性偏正结构的中心语等，而名词“期间”只具备上述功能的最后一项。（3）不同词类的词，其语法功能也许反而相似。例如，形容词“富裕”跟动词“信任”，形容词“虚假”跟名词“实物”等等。（4）一些词的语法功能没有得到充分的描写。例如，“期间”用在体词性偏正结构中，另一直接成分通常是动词或动词性结构，把“期间”看作名词或现有词类体系中的其他词类都不太合适。（5）缺乏对词的各种语法功能的定量描写。例如，一个词能作主宾语的概率是多少，

① 邢福义：《汉语语法学》，东北师范大学出版社 1996 年版，第 294 页。

作谓语的概率是多少，这种数据对于自动句法分析很有用处。[①] 目前还只有对词类语法功能频率的一些小规模调查。这样粗糙的词类体系在句法分析中能起到多大作用呢？我们不妨举个例子来看一下。例如“接待/v 两/m 位/q 领导/n 期间/n”这个短语，计算机在进行自动句法分析时，处理的是这样一个词类标记序列：

VT　M　Q　N　N

我们希望给计算机提供一套形如“NP + VP→S”的句法规则后能够得到正确的句法分析结果，实际上这样的句法分析难度是很大的。任何一位读者，如果仅仅知道 VT、M、Q、N 分别代表及物动词、数词、量词、名词，不看具体的词语序列，都很难确定标记序列所对应的是哪一种句法结构，更何况是机器。对于这样的情况，计算机只能给出所有可能的句法结构。对于上述例子，在人看来是没有句法歧义的，因为人看到词语序列中的每个词语时都能激活跟具体词语相联系的许多知识，而在计算机看来却是充满歧义的。怎样才能使计算机在分析的过程中自动选择正确的结构分析，那就只能尽量为计算机所面对的每一个词类标记提供尽可能详细的信息。就现有的词类体系看，“领导”和“期间”的语法功能差别很大却给以相同的标记，这样粗糙的词类知识，难以有效地支持自动句法分析。诚然，自动句法分析中的歧义现象并非都是由语法方面的因素造成的，还有语义等其他方面的一些因素，所以我们构建的实词词类体系，并不奢望能解决自动句法分析中碰到的由于复杂语义和篇章层面等因素造成的诸多问题，只是期望对因句法关系的不同而产生的歧义能起到一定的作用。

在整个研究过程中，面对上述第一个问题，促使笔者关注这项研究的实用价值，而对第二个问题的思考，则引导笔者从计算机的角度来对现有的现代汉语语法理论和具体的语言研究工作进行评

① 陈小荷：“从自动句法分析角度看现代汉语词类问题”，《语言教学与研究》1999 年第 3 期。

估，进而自觉地追求本课题研究所希望达到的、对自动句法分析和现代汉语语法理论建设有所贡献的目标。

二 我们的立场和主张

在具体论述本课题研究工作之前，还有必要表明我们的理论立场。

就汉语句法描写而言，朱德熙先生的词组本位语法体系所建立的理论框架和在这个框架下开展的具体研究积累起来的成果，无疑可以看作是目前的“巨人之肩”，我们将以此为起点开始本课题的研究工作。在具体的操作过程中，我们采用的策略是导师陈小荷教授在“从自动句法分析角度看汉语词类问题”一文中提出的方法，即用句法结构作为实词归类的测试环境，主张彻底按照词充当句法成分的功能来划分汉语词类。尽管说语言学家提到的语法功能有多种类型，例如跟别的词的结合能力、连接作用、拟声作用、指代功能、充当句法成分的功能等等。其中只有充当句法成分的功能是实词分类的基本依据，其他语法功能一般是用来划分虚词类别的。汉语词类划分的问题主要集中在实词。事实上，自动句法分析时问题最多的也是实词，因为实词数量大，而且不像虚词那样几乎每一个词的语法功能都已经有详细的研究，因此我们主张用且仅用充当句法成分的功能来对实词进行分类。这一方法改变了过去为了把某词放到合适的词类中而找“特点”的做法，而是按照词的句法分布，详细描述每个词的句法功能，按照句法功能完全相同即为一类的原则对汉语中的实词进行分类。

说到句法分布，这里要做一点说明。我们所说的句法分布是指总体语法分布，即一个词所能占据的语法位置的总和，这跟有些学者主张采用部分分布观是有区别的，而实际上，现存的词类体系所依据的语法功能标准基本上都是部分分布观，即主要语法功能来分类的。我们不主张这种分类标准主要基于如下两方面的考虑：

（1）一个词类的总体功能概率与其中部分词的功能概率可以