汉语复句书读前后
语言片段的非分句识别

李　琼著

长江出版传媒　长江文艺出版社

教育部人文社会科学研究项目“汉语复句书读前后语言片段的非分句识别”
（项目编号09YJC740032）

汉语复句书读前后
语言片段的非分句识别

李　琼著

长江出版传媒
长江文艺出版社

新出图证（鄂）字 03 号
图书在版编目（CIP）数据
汉语复句书读前后语言片段的非分句识别 / 李琼 著
武汉：长江文艺出版社，2013.10

ISBN 978—7—5354—6821—5

Ⅰ.汉… Ⅱ.李… Ⅲ.汉语—复句—机器识别—研究 Ⅳ.H146.3—39

中国版本图书馆 CIP 数据核字（2013）第 154876 号

责任编辑：方 莹 责任校对：陈 琪
封面设计：川 上 责任印制：左 怡 邱 莉

出版：长江出版传媒 长江文艺出版社
地址：武汉市雄楚大街 268 号 邮编：430070
发行：长江文艺出版社
电话：027—87679360
http://www.cjlap.com
印刷：湖北新新城际数字出版印刷技术有限公司

开本：700 毫米×1000 毫米 1/16 印张：13.5
版次：2013 年 10 月第 1 版 2013 年 10 月第 1 次印刷
字数：230 千字

定价：28.00 元

内容提要

为了满足中文信息处理事业的需要，在字处理和词处理阶段取得了初步成效后，句处理工作已经提上了重要的“议事日程”，因为自然语言理解归根结底还是对语言中一个一个句子的理解。汉语的句子包括单句和复句，其中复句的机器理解又是重难点所在；除了因为复句的机器理解必然要建立在单句理解的基础之上，还有一个重要的原因是复句的机器理解牵涉到分句之间层次和逻辑语义关系的划分。同时，随着计算机软硬件技术的提高，作为对基于规则的理性主义方法的一种补充，基于统计或实例的语料库方法日益得到了计算语言学家们的青睐。正是在这种背景下，我们尝试建立一个“精加工”的现代汉语复句语料库，以期为复句的计算机理解提供相关语言知识和统计数据。

本文只是这项语料库建设工作的一部分，主要目标是在进行语料库中复句语料的分句层次和关系划分以前，首先排除掉那些不参加复句层次和关系划分的书读前后语言片段，即对非分句语言片段进行识别和标注。主要内容如下：

第一章首先回顾了前人对单复句纠结问题的研究，分析了单复句纠结的复杂现象，并尝试从认知语言学的“原型”角度看待这一客观事实，用“小句中枢”理论界定非分句（分句）的性质和范围；接着以标点符号为标记让计算机对分句进行了初步识别，排除了其中的一些非分句；最后对计算机不能以标点符号为标记识别的非分句进行了细致的分类描述，有汉语断句的随意性造成的非分句，有由于分句的某个组成成分比较复杂而促成的非分句，还有句子的某

些特殊成分单独充当一个语言片段而形成的非分句。

第二章首先介绍了词性和非分句的标注说明；接着根据“小句核心词”和“动词中心说”理论，利用标注的词性信息对一部分不包含动词的非分句进行了自动识别，并制定了一系列规则对这部分非分句实现类型的自动标注；最后，本章把一些由形式相对固定的短语充当的语言片段单独放入短语库中，通过制定一系列的短语库规则对它们进行自动识别。

第三章利用句法信息实现了一部分非分句的自动识别和标注。首先简单描述了计算机处理自然语言的工作模式；接着讨论了两种类型的“形式标记”及其对识别和标注非分句的作用：一类是显性形式标记，如介词充当的开头标记，时间词、方位词等充当的结尾标记，另一类是结构助词“的”和表判断义的动词“是”；并在此基础上制定了计算机自动识别和标注非分句的另外一系列规则，添加到规则库中。

第四章是基于前两章所编规则的实验及其结果分析。首先建立一个 access 数据库，把规则中出现的开头标记和结尾标记逐一录入到这个库中。在判断某个语言片段是否为非分句时采用简单的字符串匹配法，把每个语言片段的开头部分或结尾部分跟数据库里录入的开头标记或结尾标记进行匹配，匹配成功的就是非分句。接着通过人工方式逐条检验了规则的识别或标注正确率，并简要分析了产生错误的原因及今后的改进策略。要特别说明的是，第二章和第三章的规则都是在训练集中制定的，因此我们还要在训练集中统计这些规则的贡献率，然后把这些规则推广应用到整个复句语料库看正确率如何，同时不断改进和完善规则。

第五章尝试综合利用句法、语义和搭配知识来识别一部分非分句，这方面的工作目前尚处于试验阶段。本章首先论述了语义知识在自然语言计算机理解中的重要性，接着介绍了国内外面向计算机研究语义的概况，并给出了本文所运用的语义理论。然后针对本项研究的具体情况阐述了工作前提，包括研究语料的选取和限定、解

决问题的思路等。接下来就围绕语义角色、语义类别和语义特征等三要素，试着为108个动词的127个义项建立起动宾语义搭配框架，并提出了判断动词后面两个名词之间有无偏正关系的十八条形式标准，还运用所建立的搭配框架分析了几个子语料库2中的实例，提出了建立“动宾搭配频率表”的设想。末尾对全章内容有个小结。

第六章根据邢福义先生的有关论著从理论上制定了一系列识别名词充当核心词的分句即名词谓语句的规则，尽管训练语料库中的实际用例甚少。这部分规则并没有来得及在程序上实现，也未曾针对实际语料进行过人工检验。

关键词：小句中枢；书读前后语言片段；非分句；自动识别；句法信息；语义知识；规则

目 录

绪 论

0.1 研究背景

计算语言学是一门植根于计算机科学、数学与语言学等多学科沃土的新兴学科。二十世纪五、六十年代最初兴起时以机器翻译为中心课题。机器翻译就是利用计算机把人类的一种自然语言翻译成另一种自然语言，实现两种语言之间的语码转换。要让计算机做到这一点，首先得让计算机理解这两种自然语言，因此初期不是建立在对自然语言理解基础上的“词对词”的简单翻译方法得不到理想的翻译效果。上世纪六十年代，人们开始研究自然语言的语法、语义和语用等问题，尝试让计算机来理解自然语言，这样就出现了“自然语言理解（NLP）[①]”的研究。到了今天，计算语言学开发的应用研究领域越来越广，包括自然语言人机接口，语音的自动识别与合成，编制语词索引，情报检索，语法的检测，以及许多需要统计分析的领域，如文本考释和风格学研究等。近十年来，随着网络技术的发展和进步，人类文明也从工业化社会进入信息化社会，海量以文字、声音、图像等形式贮存的信息如潮水般向我们涌来；如何从铺天盖地的信息网络中找到自己想要的信息，这一任务迫使人们越来越重视语言信息处理技术的研究。而对我们使用汉语的人来说，中文信息的自动处理又是任务最迫切、难度最大的一项课题。

---

① 在计算机科学中，除了“自然语言理解”，也经常使用“自然语言处理”、“语言信息处理”等意义相近的术语。本文则使用“自然语言处理”。

简单地说，中文信息处理（CIP）就是利用计算机对汉语信息（包括口头的和书面的）进行处理，也就是如何让计算机自动理解和生成人类的一种自然语言——汉语。汉语的信息处理单位可以划分为字、词、短语、句子及篇章，因而学者们一般把中文信息处理的进程分为三个阶段：字处理阶段、词处理阶段和句处理阶段。

我们都知道，计算机天生是为英语等印欧系语言设计的，而汉语书面语言是用汉字字符表示的，所以中文信息处理首先面对和必须解决的难题就是汉字的输入和输出问题。目前，字处理已经取得了突破性进展：发布了汉字编码字符集国家标准，研制了国际标准ISO10646，还确定了汉字内部码；键盘输入技术、汉字识别技术、语音识别技术、汉字字形技术、语音合成技术等各项研究成果都已投入实际应用。我们还知道，汉语不像西方语言那样，词和词之间在书写的时候会留一个空格，这个空格就是天然的分词标记。汉字并排连写，词和词之间没有明显的界限，这种情况给汉语的自动句法语义分析造成了极大的困难，因此在中文信息处理的词处理阶段第一步要做的就是自动分词。1992 年国家制定了 GB/T13715－92《信息处理用现代汉语分词规范》，近几年国内的孙茂松团队还建立了一个“信息处理用现代汉语分词词表”（孙茂松 2001）。中科院、北京大学、清华大学和山西大学先后研制开发了现代汉语自动分词和词性标注系统，正确率达 95％以上，而且技术正日趋走向成熟。目前最大的难题是未登录词的识别和交集型歧义语言片段的划分。中文信息处理事业正在向句处理阶段艰难迈进。这一阶段的主要目标是“怎么让计算机处理、理解自然语言中一个句子的意义，怎么让计算机生成一个符合自然语义规则的句子”（陆俭明 2001）。为了实现这一目标，计算机需要被输入汉语的句法、语义、语用等多种知识，其中语义的知识表示和理解被视为中文信息处理事业的瓶颈。国内外很多高等院校及科研机构都在集中精力攻克句处理的难关，运用中心词驱动的短语结构语法（HPSG）、功能合一文法（FUG）、词汇功能文法（LFG）、依存语法及配价语法设计了各种句法分析

器，就句法和语义自动分析的诸多问题进行了深入的研究和探讨，力求把这些知识转换成计算机可读的形式语言。如《基于短语结构语法的自动句法分析方法》（冯志伟 2000）、《基于双语模型的汉语句法分析知识自动获取》（吕雅娟 2003）、《图算法句法分析器自动生成》（孙明勇 2003）等。

我们认为，就目前的研究形势而言，还可以把句处理阶段向下细分出短语处理、向上分出复句处理阶段。学者们以位于词和句之间的语言单位——短语作为研究的突破口，为汉语各种类型的短语设计制定了一系列自动识别和标注的规则，试图在此基础上最终实现句法的自动分析。如《汉语短语的自动划分和标注》（周强 1997）、《面向中文信息处理的现代汉语短语结构规则研究》（詹卫东 2000）、《汉语基本短语的自动识别》（张昱琪 2002）等。另外一部分研究者已经开始在大于小句的单位——复句上“动脑筋”了，这些研究旨在通过复句中关系标记的作用得到复句层次和层次关系的自动标注结果，以期在更大的单位上实现汉语篇章甚至真实文本的自动标注。如《汉英机器翻译中描述型复句的关系识别与处理》（鲁松、宋柔 2001）、《汉语复句本体模型初探》（胡金柱 2005）、《本体论在复句领域概念建模中的应用》（胡金柱 2006）。但这方面的研究工作还很少有人涉及，取得的研究成果也不多。

另一方面，随着计算机软硬件技术的提高，计算机的运算速度和存储容量有了巨大增长，为处理大规模真实文本准备了充分的物质条件；这就为自然语言处理方法的改进提供了可能性，在基于规则的技术中引入了语料库方法，其中包括统计方法、基于实例的方法等等。由此发展起来的一门新兴学科——语料库语言学（corpus linguistics）试图从大规模真实文本的语料库中获取语言知识，以求得对于自然语言规律的更为客观的、准确的认识（冯志伟 2001）。当前语料库的建设和语料库语言学的崛起，正是计算语言学战略目标逐渐转移到大规模真实文本的一个重要标志，是计算语言学研究方法上的一场革命。随着人们对大规模真实文本处理的日益关注，越

来越多的学者认识到，基于语料库的分析方法（即经验主义的方法）至少是对基于规则的理性主义分析方法的一个重要补充。如果能把这两种方法结合起来，以基于经验的统计方法作为获取知识的基本途径，并结合一些语言学家概括总结的正确的语言规则，使两者互相补充、相得益彰；那么，我们一定能够在自然语言处理领域取得较大的进展。

本文正是在上述背景下开展研究工作的：在大规模语料库中进行复句层次关系的自动分析，以期为“句处理”阶段的工作提供丰富的语言知识。

0.2 选题的意义

语料库在语言研究中的重要性已越来越为研究者所重视。语料库是最理想的语言知识资源，因为它不仅是大量语言资料的集合，还可以提供足够多的真实语料和例证，满足语言研究的需要。从知识挖掘的深度出发，我们可以把语料库分为“生语料库”和“熟语料库”两类。生语料库，又叫原始语料库（raw corpora），是没有经过任何处理的语料集合。熟语料库，又叫附码语料库（annotated corpora），是做了分词和属性标注的语料库。从标注的属性数量和加工的深度出发，熟语料库又可分为“粗加工”、“深加工”和“精加工”等不同级别。其中“粗加工语料库”是只做了分词标记的语料库，“深加工语料库”是标注了词性属性的语料库，标注了“语义”、“语用”等属性的是“精加工语料库”。所谓语料加工主要指信息属性的标注，就是把语料所具有的重要的语言学信息用字母符号代码逐项标注出来。标注得越详细，或者加工的深度越深，加工的颗粒度越小，语料库的功能和应用价值就越大。要想使语料库名符其实地成为自然语言的知识库，就必须首先对语料库中的语料进行自动标注，而且标注的属性数量越多，加工的深度越高，人们从中可以提取的语言知识就越丰富，这样的语料库价值也就越大，“语料库的

价值取决于标注的深度与准确性”（俞士汶 1993）。反过来看，一个语料库不管其规模有多大，如果收藏的仅仅是未经加工的“生”语料（即原始文本），那么它的研究价值也是很有限的。因此，为了从语料库中尽可能多地获取我们所需的各种语言知识和统计数据，必须对语料进行从词法、句法到语义、语用等多层次的加工，使生语料逐级变为精加工的熟语料。同时我们也应看到，语料库的建设是一项系统工程，并非一朝一夕能够完成。

冯志伟（2006）论述了当前自然语言处理发展的特点：基于句法—语义规则的理性主义方法受到质疑，随着语料库建设和语料库语言学的崛起，大规模真实文本的处理成为自然语言处理的主要战略目标。而事实上，当面对铺天盖地向我们涌来的大规模真实文本时，我们发现复句的数量远远大于单句。复句是由两个或两个以上的单句联结形成的，其信息容量比单句大，往往能表达比单句更复杂的语义内涵。作为一种特殊的语法实体，复句上连篇章，下连小句，在篇章和小句之间搭建了一座沟通的桥梁，同时具有语义、语法和语用等多方面的属性。我们的目标正是建立一个面向中文信息处理的现代汉语复句语料库，这个语料库必须是经过“精加工”的，可以为机器翻译、信息检索、自动索引和文摘等研究工作提供相关的语言知识和统计数据，充分发挥语料库语言学和统计数学方法在中文信息处理事业中的重要作用。例如，要想让计算机正确地把下列汉语复句翻译成英文，我们必须提供给计算机的知识有很多，其中重要的一项便是“复句的层次关系”；如果计算机不能厘清分句之间的层次关系，机器翻译必然不会取得理想的效果。

（1）他的性格，在我的眼里和心里是伟大的，虽然他的姓名并不为许多人所知道。（鲁迅《藤野先生》转引自《新实用汉译英教程》陈宏薇编著湖北教育出版社 2000 年版）

（2）In my eyes and heart he is a great man, though his name is not known to many people.（译文略有改动）

就上面这个复句来说，计算机必须要有“‘他的性格’不是一个

分句”、“两个分句之间是转折关系”等语言知识，才能得到相对正确的翻译结果。计算机如何获得这些语言知识？其中一个重要的途径或来源便是语料库，因为语料库的主要作用就是为统计语言模型提供语言特征信息和概率数据。也就是说，如果我们能建立起一个大规模的复句“精加工”语料库，实现复句层次关系的正确标注，将为复句的机器翻译研究提供一种基于统计和实例的分析技术，与那种基于规则和研究者个人语感的分析技术形成优势互补的良好局面。

我们把这个语料库建设和应用的过程图示如下：

要建立这样一个面向应用的大规模现代汉语复句语料库，我们必须完成如下三大部分工作：一、自动分词和词性标注，我们目前用的是中科院计算语言所开发设计的自动分词软件 FreeICTCLAS；二、小句[①]（单句、分句）句法和语义结构的自动分析，其中包括短语的边界识别和标注；三、分句间层次和关系的自动切分和标注。目前自动分词和词性标注工作已暂告段落[②]，小句的自动分析也正在如火如荼的进行中，现在要着手开展第三部分的工作了。

然而，由于汉语缺乏严格意义上的形态变化，断句比较随意，在真实的文本中，小句和小句之间甚至小句的成分和成分之间经常呈现似断似连的局面。特别是由于强调、语气等语用因素，或在某种特殊情况下出于交际的临时需要，本来一句话可以一气呵成，却偏偏分成几段说出来，中间被停顿或语气词隔开；在书面语中就表现为，明明连着的一个完整小句，中间可能会被标点符号隔开。从

① 本文的小句取邢福义先生《汉语语法学》中的定义：“小句”主要指单句，也包括结构上相当于或大体相当于单句的分句。

② 词性代码表见附录一。

这一点来看，我们可以说汉语是最忠实于思想的语言，“想到哪，说到哪”；我们也可以说汉语是最没有“组织纪律性”的语言，毫无章法可言，有时甚至是杂乱无章的。例如：

（3）我，数学考试，得了第一名。

（4）10 年来，我们的决策在民主化和科学化方面，已经有了很大进展，否则无法解释 10 年来我们取得的伟大成就。

为了论述的方便，在这里我们首先要定义一个概念——书读前后语言片段[①]（下文简称语言片段）。“书读”即书面上的标点 号，它前面或后面的那部分就是一个语言片段，可能是一个词，可能是一个短语，也可能是一个小句，还有可能是“四不像”。如例（3）有三个语言片段，例（4）有四个语言片段。但例（3）实际上是一个小句，说话人或作者出于某种语用的需要，或者是强调，或者是心情兴奋，把一个完整的小句拆分成三个部分说出来，书面上用两个逗号隔开成三个语言片段。例（4）是一个复句，表面看来有四个语言片段，其实只包含两个分句；因为前两个语言片段都不是复句的分句，只是由于表达的需要被标点符号隔开的分句的一部分（状语成分）。

在我们要建设的现代汉语复句语料库中就有很多类似于例（4）的复句，这些复句包含非分句语言片段，而这些语言片段又会影响和干扰复句层次关系自动切分和标注的结果，降低其准确率。因为考虑到编程的方便，而标点符号又是最直观的标记，计算机在处理复句语料时最初是以标点符号作为分句标志的。也就是说，计算机会自动把每个用标点符号隔开的部分标注为分句，然后再划分它们之间的层次并标注其关系。还是拿例（4）来说，计算机在处理这个复句时，会以标点符号为划分标记把它分为四个分句；而其实第一和第二个语言片段都不是分句，不应该参与到复句层次关系的划分和标记，否则会使层次划分进而使关系标注出现错误。因此，在进

① 这个概念是在和邢福义先生、汪国胜先生的谈话中由邢先生首先提出来的。

行复句层次关系自动划分和标注的研究工作之前，应首先排除那些不是分句的语言片段。

由此可见，本文的研究工作实际上只是整个“复句信息工程”的一个子项目——书读前后语言片段的非分句识别和标注，这也是整个“复句信息工程”中一项重要的、基础性的工作。

本项研究所用语料库是由华中师范大学语言研究所主持开发的现代汉语复句语料库。该语料库收有65万个复句，语料主要来源于《人民日报》和《长江日报》，目前已完成了自动分词和词性标注工作，复句关系词语的识别和标注工作也已取得阶段性成果，下一阶段就要进行复句层次和关系的自动划分和标注了，而在此之前必须进行的本项研究——书读前后语言片段的非分句识别和标注工作将为此打下坚实的基础，为今后正确地划分和标注复句层次关系提供基本保证。从这个角度来看，本项研究是属于自然语言理解领域的。另一方面，冯志伟在《计算语言学基础》一书中对语料库语言学进行了详细、准确的定义：语料库语言学主要研究机器可读自然文本的采集、存储、检索、统计、语法标注、句法语义分析，以及具有上述功能的语料库在语言定量分析、词典编纂、作品风格分析、自然语言理解和机器翻译等领域中的应用。因而从这个角度来说，本项研究工作又应纳入语料库语言学的研究范围。

0.3 本文的理论背景

“小句中枢”理论认为，小句是最小的具有表述性和独立性的语法单位，主要指单句，也包括结构上相当于或大体相当于单句的分句。小句具有表述性，能够表明说话的一个意旨，体现一个特定的意图；小句具有独立性，一个小句不被包含在另一个小句之中；小句还是最小的语法单位，每个小句都带有特定的语气。

在汉语的各级各类语法实体中，小句处于中枢的地位，所具备的语法因素最为齐全；内联词和短语，外联复句和句群，还同一定