国家社会科学基金重大项目研究成果
新闻出版署“十五”规划重点图书

中文信息处理

ZHONGWEN XINXI CHULI

现代汉语词汇研究

XIANDAI HANYU CIHUI YANJIU

许嘉璐　傅永和 \ 主　编

广东省出版集团
广东教育出版社

国家社会科学基金重大项目研究成果 ⊙ 新闻出版署“十五”规划重点图书

中文信息处理

ZHONGWEN XINXI CHULI

XIANDAI HANYU CIHUI YANJIU

现代汉语词汇研究

许嘉璐　傅永和 \ 主　编

广东省出版集团
广东教育出版社

图书在版编目（CIP）数据

中文信息处理现代汉语词汇研究/许嘉璐，傅永和主编.—广州：广东教育出版社，2006.9
ISBN 7-5406-6300-6

Ⅰ.中… Ⅱ.①许… ②傅… Ⅲ.汉字信息处理-词汇-研究 Ⅳ.TP391.12-61

中国版本图书馆 CIP 数据核字（2006）第 031744 号

广东教育出版社出版发行
（广州市环市东路 472 号 12-15 楼）
邮政编码：510075
网址：http://www.gjs.cn
广东新华发行集团股份有限公司经销
佛山市浩文彩色印刷有限公司印刷
（南海区狮山科技工业园A区）
787 毫米×1092 毫米 16 开本 36.25 印张 610 000 字
2006 年 9 月第 1 版 2006 年 9 月第 1 次印刷
印数 1-3 000 册
ISBN 7-5406-6300-6/TP·16
定价：68.00 元
质量监督电话：020-87613102 购书咨询电话：020-34120440

序

在当前计算机极快地普及，互联网通往千楼万宇，各种信息以疯狂的速度增长的时代，人们越来越感到利用计算机处理语言文字的重要性，也越来越迫切地期望计算机能够具有高水平的语言文字的能力。但是由于汉语的特点，计算机在处理汉语信息（包括文本语言——书面语、自然语言——口语）时，遇到了许许多多难题，以至于到现在还没有突破性的进展。这在计算机用户的直觉中就是写作时输入的只是词，检索时也是靠词语，因而很不准确；要对文献进行自动摘要，还不可能；想把浩如烟海的信息自动地加以分类，常常张冠李戴；至于想把中文翻译成英文或其他外语，译文竟能让人哭笑不得……总之，要“自动”就不准确，要准确只好靠手工。

上面所列举的人们的直觉，翻过来看就是对计算机进行中文信息自动化处理的期待。

为什么眼前的计算机不能满足人们的这些需求？不是因为硬件不能承担对汉语及其书写符号汉字进行“计算”的任务，而是软件没有给计算机提供足够的计算“内容”——有关语言的知识。而这又是因为我们研究计算机和汉语的人们还没有掌握足够的有关现代汉语的规律。

这个结论似乎与人们以往对现代汉语界研究情况的印象不一样。这是因为，过去的研究是为人服务的，如果以帮助人们深入认识自己所使用的语言，应该说近百年来学者们的努力已经给社会提供了相当丰富的知识；但这些知识不是为计算机服务的，如果以计算机“理解”语言所需要的知识为标准来衡量，则到现在为止所取得的研究成果，还远远不够。

人和计算机最主要的区别在于人有语感，机器没有。语感，即人对语言所表达的意义与语言环境（包括主观客观、显性隐性的种种语言之外的因素）的关系的一种不言而喻的直接感觉，这种感觉可以大大补充语言表面所未能表达的意思和情感。语感，是人在学习和使用语言过程中，与使用同一语言

的人达成的默契，是民族传统影响下形成的心理能力。人因为有语感，定性式的知识就可以满足人的相当程度的需要。人还有逻辑类推的能力，给一个定义，几个例证，人就可以领会、掌握、使用、举一反三。而计算机，是按照人所提供的条件（知识）进行工作（计算）的，条件不足，计算机就不会类推。计算机所需要的条件，是尽量充足的与语言有关的知识，不但有语言的规则，而且有语言之外的知识（有人称之为“世界知识”，我认为这一术语还不能完全准确地揭示语言之外知识的全部本质，姑且不用）。因此，教给计算机必要的语言规则、对语言环境的分析和“语感”，就成了解决中文信息自动化处理问题的必经之途。

教给计算机有关语言的知识，自然不能完全脱离人们已经形成的对语言的认识；也就是说，应该充分吸收已有的研究成果，并在其基础上朝着计算机所需要的方面发展。

正是基于以上的考虑，我和一批从事现代汉语研究和中文信息处理研究的著名专家齐心合力地承担了国家社会科学“九五”重大项目“面向计算机的现代汉语词汇研究”。经大家几年的努力，课题圆满地结束了。

在结题的专家验收会上，验收组和社会科学规划办公室都指出，课题组最好能把研究的成果结集成书，供全社会参考。这是个好主意。但是主持各个子课题的专家工作都极其繁忙，这个课题一结束就又投身于下一步的研究了。时至2004年，我与同仁才找到了可以坐到一起研究写作的时间。商量的结论是马上合力撰写这本《中文信息处理现代汉语词汇研究》。

这本书可以有三种编写法。一种是每个课题写一篇，编成论文集；一种是按照项目内容原本就存在的内在联系写成一本概论式的专著；一种是编成专著，但照顾到各个子课题的特点，既相互衔接，各篇又相对独立。我们采取了第三种方案。这是因为，我们考虑到读者中不乏有志于中文信息处理技术的年轻人，比较系统和有序地安排，“把珍珠串成项链”，适合他们全面了解这一领域的动态和存在的难点，因而可以作为有关学科本科或研究生的参考书；同时，又可以突出各篇的独立性，减少统稿的工作量。现在呈献给读者的就是这样一本书。

本书的每一章，都由各个子课题的负责人亲自撰写，由傅永和先生和我分别审阅，提出修改意见，退作者改定后交出版社。

我们既然想到这本书可能被有些教师和同学作为教学和学习参考，所以几乎每章都有关于这一专题的基本知识和研究者何以采取这样的策略的叙

述，有的还如实介绍了思考的过程，而不仅仅是写出结论。我们戏称这是“肯把金针度与人”。

在这里，我想交代一下这一项目进行的情况，或许对读者阅读此书和开展相关研究有所参考。课题研究开始前，我们先进行了认真的长时间的论证。专家们从国内外中文信息处理研究的现状、国内技术力量分布、词汇研究所不可缺少的诸方面条件等角度反复讨论；有关专家还展示了自己研究的成果。论证的结论是：必须尽快地加强中文信息处理研究，占领这一领域的制高点；面向中文信息处理的现代汉语研究，从词汇入手是对的；词的切分、词类划分、语料库和知识库建设应该同时进行；各个子课题间应该互补互学，资源共享，分而有合，整个项目应该是一个整体。

我们这个项目的核心既然是词汇研究，所以子课题的设计不但围绕着计算机所需要的词汇知识，而且尽量彼此衔接，即课题之间有一定的逻辑关系；同时，我们不是为研究而研究，最终结果不应该只是几篇论文，还要有可以供更多的人使用的软件，包括数据库。

概括言之，在这一项目中，要解决的主要问题是：词的切分、消歧，词性的确定和标注，词与词的关系，汇总已有知识和成果的电子词典。这些主要问题呈阶梯型关系，所谓“逻辑关系”，即指此。

我们同时决定，各个子课题除了及时报告研究进行情况、项目主持人不时了解各子课题的进展外，全组还要举行一些学术研讨会，并通过种种措施在出成果的同时也出人才，即培养年轻的学术技术骨干。各个子课题组此前都已经有了相当丰厚的学术与技术积累，在本项目中所承担的子课题，也都是在各自原有研究的基础上进行的。不止一个子课题进入这个项目时正面临着只要再推一把，就可以出现较大进展的局面。既然不是一切从零开始，各个子课题间自然就不能衔接得天衣无缝，而且彼此一点都不交叉。这一点或许是目前不同单位的人结合在一起开展集体研究所难以避免的，因此在事前论证时，大家一致认为，这种在一定程度上的重合是允许的。例如，关于词的切分、词性标注等方面就是如此。开展这一课题研究的三年多，全组即按照论证时的设想进行，应该说，预定的目标全部达到了。

大家都知道，在我国进行中文信息处理研究的过程中，有不少工具性的研究和建设常常重复，例如语料库。人们指出，这是很大的浪费。的确如此。在我们对课题进行论证时就已经指出，至少在课题组内部一定要避免重复建设。但是在各个子课题开始之后，却发现要做到这一点是非常困难的。这是

因为，各位专家的研究并不是从零开始，此前已经有了相当的基础，包括工具的准备；各位专家所依据或创造的理论和研究思路不同，从而导致设计的差异。“生米已经做成半熟饭”，重新设计和建设一套工具费时费力，且为我们所领到的经费所不能支。另外，按照理论各个子课题间可以彼此借助的，由于是同时开始、并行前进的，也没能实现某些资源共享的理想。由此我想到，在中文信息处理领域，在一个时期或阶段有所重复几乎是不能完全避免的，这也是我国研究水平还不够高的表现；同时，必要的重复或许可以造成或促成流派的形成。因此，读者在本书的不同章节中可以看到不同子课题关于各自语料库、知识库建设的介绍，其中就有应该统一而不能统一的问题。例如关于词的切分、词性标注、消歧，等等。

那么，到现在为止，我们做了哪些工作，距离计算机对语言进行自动化处理还有多远的路程呢？

为中文信息处理服务的现代汉语词汇研究，首先要解决什么是“词”。过去语法学家给出的关于“词”的定义，需要拿到足够数量的词汇中进行检验、修订和补充，以确定“词”的界限。但是，这里所说的“词”，和传统意义有所不同。笼统地说，即凡是只能连在一起使用，一般人的语感把它当作一个单位的两个或更多的字符，也作为词对待（通常称为“切分单位”）。“词”界定了，接着就要把文本语言和自然语言线型的字符串以词为单位切分开来。这是汉语特有的问题和工序。用汉字写或印出的东西，字和字之间是等距的，计算机不能像“读”拼音文字那样根据词与词间的空格轻易地分辨出词，因而需要教给它如何把字符串以词为单位分开。有了上面这两项，计算机就可以将文本文件比较准确地分出一个个词。

语言，无论是文本语言还是自然语言，都是成段、成片地存在的。要对语言进行自动化处理，计算机只“认识”词，却不能理解成串的话，这还远远不够。因此，接下来就需要把每个词的语法性质、意义、和别的词之间可能发生的关系一一标注出来，以便计算机能够“算”出成段、成片语言的语法结构、每个词的功能和意义。这是让计算机能把文本文件作为一个整体来理解的基础。

但是，我们这个项目仅仅是面向计算机的现代汉语词汇研究，其根本性的任务是设法让计算机“了解”、“掌握”词汇；至于语言中大于词汇的单位和超出语法范围的问题，都不能在这个项目中解决。

在现实语言生活中，语言是极为复杂的思维活动的外在和物化（语音和

文字的形体都是物质的)。人们说话和写作时会根据说写的目的而运用不同的文体，还因为语言环境的影响而错综其事、藻饰其词、繁简相得，如果想让计算机对这些捉摸不定的情况也能处理，还要进行大量的研究。只有计算机对语言的理解不限于形式和表层的意义时，语言的自动化处理才接近人们的需要和期待。如果用中文信息处理的长远目标来衡量，目前我们所做的工作不过是其小焉者，充其量只是长征开始时的一部分准备工作而已。事实上，这个项目的成果多数是今后做进一步研究时所需要的工具。例如电子词典、词的切分软件、规范词表等即是。如何使用、改进这些工具以进行更加深入的研究，还需要人们继续思考和探索。

科学研究应该具有“进攻性”——向未知领域挑战，“明知山有虎，偏向虎山行”；而要具有“进攻性”，就必须具备预测性——对科学发展前景的估计（依据科学现状及其发展的逻辑)。就中文信息处理而言，进攻性应该体现为，在初步解决了词汇问题之后，朝着句、段、篇进军；预测性则应该是准确地判断难点之所在，在试验着寻找解决方案时，预估其可行性和成功的概率。

正是从科学研究的这两性出发，我在这一项目结束后又大胆地在“863”计划中申请了一个项目——“面向计算机的现代汉语研究”。由“词汇”到“现代汉语”整体，其继承性和延伸性显而易见。我之所以说大胆，是因为虽然申请之前依然请许多中文信息处理专家进行了论证，大家一致认为项目可以取得理想的效果，各个子课题的学术带头人也充满信心（其中有好几位就是社科“九五”重大项目的成员)，但作为主持人，我心里其实并不踏实：科学研究毕竟是成功与失败并存，现有的理论虽然都言之成理，能够自圆其说，但是能不能在技术上实现，能不能走出实验室应用于更广大的领域，难以预知；何况，在社会上，“科学研究允许失败”的道理和实际情况差距很大，如果项目不成功，共同攻关的学者们能不能坦然面对种种舆论，继续坚持下去？

还好，项目进行近三年了，经过学者们的刻苦工作，无论是应用基础研究还是应用研究，大部分子课题都取得了预期结果，有的还超出了事先的设计和预期。例如“电子词典”、“现代汉语词典自动查重与校对系统”、“基于语境和立场判断的检索”等都已经达到了实用水平。

如果与这本书所反映的成果相比，“863”项目的最大特点是以语义为核心开展研究。这比以语法为重前进了一大步。语义问题是中文信息处理的核

心，在理论建设与技术实现方面也是瓶颈，在我，这是经过十年的探索获得的越来越深刻的认识。在中文信息处理界，这可能是早已取得的共识。

我们确定了三个流派携手并进的原则，在电子词典、词与词的关系、直接切入语义等方面同时展开，一旦在哪一点上有所突破，便及时把技术转化为产品，推广使用，既服务社会，同时还可以获得一些经济效益，支持项目的开展，而技术与理论也在使用中得到检验，有利于其逐步完善。

研究语义，《现代汉语词典》等优秀辞书固然为计算机的语义知识库准备了较好的基础，但是，与语义密切相关的语境、语体、风格等——其中很大成分上属于民族文化问题——只靠词性、词的义项、词的语法功能还是束手无策的。看来需要改变一下习惯性思维，多几条思路，尽力体现人在学习、理解语言时的心理过程，或许可以找到解决的办法。其中，我国古老的学科训诂学，可以给我们很多有益的启发。也可以说，“经验主义”的作用在这一领域可能并不比“理性主义”的作用小，关键是正确地总结出人类掌握语言的“经验”。在我们每个人已经熟练地掌握了某种语言后，起始阶段的心理过程就已经“忘记”了，否则现在研究起来要容易得多了。在这方面，现代实验心理学似乎也帮不上多大的忙，主要还是要靠计算机专家和语言学家的紧密合作，需要长时间的努力，令人高兴的是，在我们这个“863”项目中，已经露出了一线曙光。

“面向计算机的现代汉语词汇研究”的成果虽然至今在国内还处于领先的地位，并且还在被人们使用着，但是在我看来毕竟已经成为过去的事。我们的眼光已经越过了它。但是我希望读者能够领会到这一项目的“历史意义”——在一定程度上可以说，没有这个项目的成功进行，就没有“面向计算机的现代汉语研究”项目的开展，就没有如今在语义研究方面的进步，从团结合作、励志奋发的团队的形成，到理论和技术的发展，以及项目实施的管理，“词汇研究”都起到打基础、铺阶梯的作用，所以现在结集出版这本书的价值已经超出了中文信息处理的理论和技术层面。

感谢广东教育出版社慨然承担了出版这本排版、校对都比较困难的书的任务。我相信，他们所获得的社会效益一定大于他们辛勤的付出。

许嘉璐

2005 年 8 月 1 日于

日读一卷书屋

CONTENTS

目 录

目 录

CONTENTS

第一章
信息处理用现代汉语分词词表

一、问题的提出及其意义

“词是什么”（词的定义）及“什么是词”（词的具体界定）一直是汉语语言学界中纠缠不清却又挥之不去的基本问题。词的定义，一般来说争议不大（虽然因看问题角度的不同，也有“语音词”、“词汇词”、“语法词”、“形式词”、“理论词”等说法），其典型表述为：“词是最小的能够独立活动的有意义的语言成分”[1]，或“词是具有语音形态，又能表示特定意义，且能在句法上单独出现或与其他的词共同形成词组的最小的单位”[2]。意思大同小异，均包含了三个要素：第一，词是有意义的语言成分；第二，词必须具有独立活动的能力；第三，对第一个要素进一步施加限制，词应是有意义的语言成分中的最小者。概念上清晰之至，但当应用此定义对具体的词进行界定时，仍会遇到相当的困难：某些情况下词和非词的界限并不好掌握，纷纷扰扰，常令人“欲说还休”，却又“欲罢不能”。麻烦出现在两头，一头是单字词与语素之间的划界（如，一般认为“讯”是不成词语素，但在新闻语料中，却常有“新华社某月某日讯”的说法，很像一个词。于是在第二个要素上产生了不一致）；另一头是复合词与词组之间的划界（如，一说“看见”

[1] 朱德熙：《语法讲义》，北京，商务印书馆，1982。

[2] 汤廷池：《汉语词法句法三集》，台北，台湾学生书局，1992。

是词，因为其中的“见”必须读轻声；另一说应是词组，因为能插入“得”或“不”，可扩展。在第三个要素上有瓜葛）。

汉语语言学研究中存在着的这种困扰，实际上是对语言社会生活的一种深刻反映。来自社会语言学的调查分析表明，以汉语为母语的普通大众关于词的语感较语言学家所规定的词的尺度要宽松得多。但公众对词的语感也有着广泛的差异，因结构而异，因词而异，因人而异，造成了来自另外一个方面的困扰。

进一步地，从语言工程的角度来看，汉语的词汇平面构成了现阶段中文信息处理应用领域（从传统的汉字识别、汉语语音识别及合成、文本自动校对、音转字，到最近若干年蓬勃发展的文本信息检索、搜索引擎、文本自动分类及过滤、文本摘要、语义 Web 等）的主要支撑平台，几乎没有什么应用技术可以游离于这个平面之外而存在。众所周知，英语文本词与词之间存在空格，所以对英语的信息处理一起步就是在词平面上的。而汉语文本是以字为基本单元的，词与词之间没有显式的分割符，起步是在字平面上。现有的中文信息处理应用技术（如中文搜索引擎），虽然几乎都使用了汉语自动分词系统，但由于自动分词系统的性能存在严重缺陷，导致性能不佳，更堪担忧的是，将难以向更高级的形态发展。全国人大副委员长、著名的汉语语言学家许嘉璐先生 2000 年曾一针见血地指出：“到目前为止，中文信息处理基本上还停留在‘字处理阶段’，也就是说计算机对汉语的‘认知’是一个字一个字地进行”，“如果我们说得‘宽宏’一些，最多可以说现在是处在‘字和词处理之间’阶段”。“中文信息处理技术虽然在有些方面有所进步，但至今还没有跨上‘语言处理’这个台阶”。这段话深刻之至。而要实现从字平面到“语言处理”这个平台的跨越，以汉语自动分词为典型代表的词平面是必由之途，同时也是目前来说最为现实的道路。阻挡我们前进的主要障碍之一，就是我们缺乏一个广为接受的、高质量的信息处理用汉语分词词表。来自汉语语言学研究以及公众语感这两方面的困扰，交织在一起，使得构造这样一个词表困难重重。其后果是显而易见的：中文信息处理应用系统缺乏一个构建于词平面之上的公共平台，小则给用户造成不必要的麻烦，相互掣肘（例如：“微软拼音输入法 3.0 版”将“猪肉”处理成词，而“智能 ABC 输入法 5.0 版”则将其视为两个单字，所以当用户使用前者输入这两个字时，一次成功，无须选择，而使用后者则需要逐字选择。这种操作上的不一致会导致使用这两个系统的同一个用户感到别扭），大则导致应用系统的发展后劲严重不足

(例如：真正意义上的语义 Web 的研发依赖于广为接受的、高质量的通用本体体系和各领域本体体系，而这些本体体系的构造离不开一个广为接受的、高质量的词表)。

这样一个词表的影响，不仅仅体现在各种应用系统中，也会作用于计算语言学的研究上。这里想以汉语分词语料库的一致性问题为例，适当展开，予以说明。

我们知道，一个经过分词处理的大型汉语语料库不仅是进行语言学研究“气韵生动”的素材库，而且也是进行计算语言学研究的宝贵资源。严格说来，在世界范围内，一个真正经得起各方面推敲的大型汉语分词语料库迄今为止还没有（虽然海内外已有若干分词语料库陆续推出，著名的如北京大学计算语言学研究所与富士通公司联合研制的《人民日报》语料库及国家语委研制的现代汉语平衡语料库)。问题的要害在于分词语料库的质量。而衡量质量的重要标准之一是分词后的语料库是否具有比较高的一致性。关于分词的一致性大致包含两方面的内容：①一致性—1：在保持语义同一性的前提下，一个成分在语料库中的分合是否始终一致（例如：“猪肉”是否始终保持一个整体，或者始终分开)；②一致性—2：与某个成分具有相同结构类型的其他一切成分在语料库中的分合是否与该成分始终一致（例如：“牛肉”与“猪肉”的结构类型完全相同，“牛肉”是否跟随了“猪肉”的分合状态)。现在我们考察一下清华大学智能技术与系统国家重点实验室与北京语言文化大学语言信息处理研究所合作研制的一个规模在 200 万字左右的、经多级加工的汉语语料库“HuaYu”。首先，我们在国家标准 GB/T 13715-92《信息处理用现代汉语分词规范》（下文简称《规范》）的基础上，经修改、补充、细化，自己拟订了一个很细致的分词规范，作为操作依据，以期有效降低标注结果中的不一致。标注者为若干名语言学教师及研究生，在标注前已认真学习了该分词规范。然后，对“HuaYu”进行了多轮人工分词的处理。我们对第一轮标注结果进行了分析、总结（因为第一轮颇能反映标注者的语感)，发现人工分词后的“HuaYu”，仍然出现了相当多的标注不一致的情况。产生不一致（这里主要针对相对简单的一致性—1 问题，还没有涉及更为复杂的一致性—2 问题）的主要结构类型包括：

(1) 定中结构

(1a) 名词+名词

【车门】(合/总分合[1]：93.3%)　【雨水】(合/总分合：60.0%)
【酒瓶】(合/总分合：50.0%)　【烟厂】(合/总分合：45.6%)
【国防部】(合/总分合：87.5%)　【中国人】(合/总分合：66.7%)
【北京人】(合/总分合：16.7%)

(1b) 形容词+名词

【蓝天】(合/总分合：88.9%)　【白发】(合/总分合：69.2%)
【绿树】(合/总分合：37.5%)　【黄土地】(合/总分合：76.2%)
【大海】(合/总分合：95.7%)　【大桥】(合/总分合：83.3%)
【大雨】(合/总分合：70.6%)　【大树】(合/总分合：41.2%)
【强国】(合/总分合：84.6%)　【真话】(合/总分合：87.5%)
【坏人】(合/总分合：70.0%)　【古城】(合/总分合：63.2%)
【好人】(合/总分合：60.0%)　【新药】(合/总分合：6.7%)

(2) 状中结构

【深知】(合/总分合：94.1%)　【改建】(合/总分合：90.9%)
【分管】(合/总分合：90.0%)　【重建】(合/总分合：78.3%)
【猛扑】(合/总分合：62.5%)　【新建】(合/总分合：6.7%)

(3) 动宾结构

【唱歌】(合/总分合：96.7%)　【吸烟】(合/总分合：81.3%)
【下雨】(合/总分合：75.9%)　【低头】(合/总分合：53.3%)
【挥手】(合/总分合：47.1%)　【做饭】(合/总分合：46.7%)
【开车】(合/总分合：30.8%)　【买饭】(合/总分合：20.0%)
【养猪】(合/总分合：17.6%)　【敲门】(合/总分合：4.2%)

(4) 动补结构

(4a) 动词+动词

【打断】(合/总分合：95.0%)　【惊醒】(合/总分合：91.3%)
【看成】(合/总分合：86.6%)　【传遍】(合/总分合：83.3%)
【去掉】(合/总分合：75.0%)　【建成】(合/总分合：66.7%)
【盯住】(合/总分合：58.8%)　【铺通】(合/总分合：50.0%)

(4b) 动词+形容词

[1]“合/总分合”指成分在分词“HuaYu”(第一轮标注结果)中取“合”的次数占其总出现次数(“分”+“合”)的比例。

【缩短】(合/总分合：95.0%)　【降低】(合/总分合：90.0%)
【减轻】(合/总分合：86.7%)　【加深】(合/总分合：66.7%)
【搞好】(合/总分合：40.0%)　【认准】(合/总分合：15.3%)

(4c) 动词+趋向动词

【露出】(合/总分合：94.4%)　【发起】(合/总分合：90.5%)
【揭开】(合/总分合：87.5%)　【穿过】{庭院}(合/总分合：76.7%)
【翻开】(合/总分合：34.8%)　【考上】(合/总分合：23.3%)
【离去】(合/总分合：28.6%)　【跨进】(合/总分合：10.0%)

(4d) 动词（形容词）+介词

【等于】(合/总分合：99.9%)　【大于】(合/总分合：95.0%)
【出于】(合/总分合：86.7%)　【小于】(合/总分合：80.0%)
【始于】(合/总分合：44.4%)　【低于】(合/总分合：21.4%)
【优于】(合/总分合：20.0%)　【高于】(合/总分合：10.0%)
【取决于】(合/总分合：56.7%)
【献给】(合/总分合：70.0%)　【留给】(合/总分合：53.3%)
【走向】(合/总分合：46.7%)　【推向】(合/总分合：3.3%)

(5) 复杂概念名词（包括各领域术语）

【多发病】合：多发病；分：多 发 病 (合/总分合：33.3%)
【上层建筑】(合/总分合：61.9%)
【社会关系】(合/总分合：23.8%)
【三中全会】合：三中全会；分：三 中 全会 (合/总分合：13.3%)

(6) 半凝固格式或习用语

【一动不动】合：一动不动；分：一 动 不 动 (合/总分合：84.6%)
【密不可分】合：密不可分；分：密 不 可 分 (合/总分合：60.0%)
【各就各位】合：各就各位；分：各 就 各 位 (合/总分合：33.3%)
【从头到尾】合：从头到尾；分：从 头 到 尾 (合/总分合：25.0%)
【真抓实干】合：真抓实干；分：真 抓 实 干；真 抓 实干 (合/总分合：23.1%)
【招商引资】合：招商引资；分：招 商 引 资；招商 引资 (合/总分合：33.3%)
【与此同时】合：与此同时；分：与 此 同时 (合/总分合：16.7%)

(7) 其他

(7a) 副词重叠

【从未】(合/总分合：90.0%)　　【不曾】(合/总分合：86.7%)

【未曾】(合/总分合：85.7%)　　【并不】(合/总分合：73.4%)

【无不】(合/总分合：54.5%)　　【从不】(合/总分合：6.7%)

(7b) 能愿动词重叠

【不得不】(合/总分合：99.9%)　　【不能不】(合/总分合：66.7%)

(7c) 缩略语

【国内外】合：国内外；分：国 内 外 (合/总分合：90.0%)

【东西方】合：东西方；分：东 西 方 (合/总分合：69.2%)

【高新技术】合：高新技术；分：高 新 技 术；高新 技 术 (合/总分合：69.2%)

【摇摇头】 合：摇摇头；分：摇 摇 头 (合/总分合：60.7%)

(7d) 加缀

【工业化】(合/总分合：90.0%)

【多样化】(合/总分合：86.7%)

【老朋友】(合/总分合：50.0%)

可见，即使是具备一定语言学训练的标注者，即使事先已经有了规定相当细致的分词标注规范作为标注的指导，标注出来的分词语料库的质量由于一致性问题，并不容易得到保障。在质量差强人意的语料库上进行种种研究，所取得结果的可靠性当然要打相当的折扣。如果我们有一个高质量且对文本覆盖能力很强的词表，并且在这个词表中，充分照顾到了一致性—2 问题，则据之所得到的分词语料库，无论是其一致性—1 问题还是一致性—2 问题，都可望得到有效的控制（至少可以保证在语料库中占绝大部分的常用成分的分词一致性，其余非常用成分的分、合则可以见仁见智，允许有一定的自由度，但已经无碍大局）。

以上的讨论显示，这项工作的意义不言而喻。

二、对策与原则

面对词的具体界定这么一个困难问题，怎样才能找到一条切实可行的出路呢？《规范》提出了一种新的策略：定义了一个新的概念“分词单位”