计算机时代的汉语和汉字研究

罗振声　袁毓林　主编

清华大学出版社

(京) 新登字 158 号

内 容 简 介

本书是“计算机时代的汉语和汉字研究学术讨论会”(1995 年 12 月 5 日—1995 年 12 月 7 日,北京)的论文集。书中收入的 56 篇论文是从大会征集论文中精选出来的。内容包括:语言学理论和汉语句法语义研究、句法和语义的自动分析、语料库研究和语料库语言学、语音识别和汉字识别、汉字分析和汉字输入技术、汉字研究和汉字教学等六个方面。

本书反映了近年来我国理论语言学和计算机语言学研究与应用的最新进展,对今后的中文信息处理的基础研究、开发和应用都具有重要的参考价值。

本书可供语言学、计算机等专业的科研人员、工程技术人员、大学教师、研究生和高年级大学生阅读。

图书在版编目(CIP)数据

计算机时代的汉语与汉字研究/罗振声,袁毓林编著. —北京:清华大学出版社,1996

ISBN 7-302-02242-9

Ⅰ. 计… Ⅱ. ①罗… ②袁… Ⅲ. 汉字信息处理-语言信息处理学-研究 Ⅳ. TP391

中国版本图书馆 CIP 数据核字(96)第 16196 号

出 版 者:清华大学出版社(北京清华大学校内,邮编 100084)
责任编辑:魏荣桥
印 刷 者:北京密云胶印厂
发 行 者:新华书店总店北京科技发行所
开 本:850×1168 1/32 印张:15 5/8 字数:404 千字
版 次:1996 年 11 月第 1 版 1996 年 11 月第 1 次印刷
书 号:ISBN 7-302-02242-9/TP·1092
印 数:0001—1500
定 价:20.00 元

第一部分

总　　论

语言学研究与中文信息处理

许嘉璐
(国家语言文字工作委员会)

我们的时代是信息化的时代,是计算机走进万千家庭、成为人们工作生活必需品的时代。中国正在出现信息现代化建设和计算机普及高潮,“八五”计划胜利完成、“九五”计划即将开始的时候,举办以“计算机时代的汉语和汉字研究”为主题的讨论会,是很有意义的。原因很简单,要使计算机成为人人可用的工具,就需要加快软件的汉化,而要使汉语汉字能顺利地在计算机中运行,就需要先研究一系列汉语汉字的规律问题。

在这次会议上,大家一定会提出很多高明的见解和问题。现在,我只想从宏观上提一点建议。

一

我们要正视当前语言学和中文信息处理技术距离很大的现实,并努力缩短这个距离。这是时代的需要,也是我国语言学界难得的机遇。语言学能否适应这一需要,能否抓住这一机遇,这不但关系到我国中文信息处理事业的发展,也关系到语言学的价值体现和是否能成为“先行科学”的问题。从语言学的整体说,既应该面向社会最急迫的需求,也应该有长远的眼光,进行扎扎实实的基础研究,而这两方面既不是彼此绝缘的,更不是相互矛盾的,而是密切相关、相互促进的。为此,语言学界的朋友们补充新知、使不同学科的交叉渗透首先在自己身上体现出来,这是时代的需要,学术

发展的需要。

二

中文信息处理领域的研究，一要强调通力协作，千万避免各行其是、彼此封锁，以致造成重复浪费、互相掣肘，国家受损，学术受阻。中文信息处理技术发展到今天，各方面能否联合与衔接，已经成了其自身今后发展的关键问题之一。二是起点要高，要避免再在低层次上投入力量，比如对基于拆分汉字的键盘输入编码方案的发明、研制、我认为可以止矣。因为由于汉字的复杂，一方面可以设计出几乎无数种编码方案，另一方面，又几乎没有一种编码方案是十全十美的。因此，旨在克服已有方案弱点的新方案可以层出不穷，同时也必然留下“破绽”再引发新的创制。这样，编码方案的增加，给国家和社会造成的损失要大于受益。现在急需要做的，是从已有的方案中选择最优者，并使之更加完善；同时，尽快在词处理、句处理、语音输入和输出等更高层次上取得突破性进展。用不了多久就可望解决的智能化汉语拼音输入方法，说不定将独霸天下，一般键盘编码输入方法则将被逐渐而迅速地、科学地、有说服力地淘汰。

三

在进行有关中文信息处理的研究时，从一开始就要注意标准化和规范化问题。语言文字的标准化和规范化，是国家统一、民族团结和社会发展的条件之一。在信息化时代，在高科技中，对标准化规范化的要求更高、更严格。语言文字的规范化标准化是国家的法律所规定的，国家的法律法规是从国家的实际情况和人民的长远利益出发而制订的。单从科学研究及其成果的经济效益考虑，也必须高度重视规范和标准，否则很可能徒劳而无功。

四

语言学工作者和计算机界的朋友,在进行前瞻性研究的同时,应该对当前的语言文字管理工作给予关心。例如社会上的用字混乱、滥用外文、胡改成语、文字表达水平普遍下降以及小学识字教学等问题,看来是浅而又浅,但是却关系到社会发展、国家的振兴,而在其背后也需要深入的科学研究作为后盾和基础。四十多年前,当国家进行文字改革和汉语规范化的时候,几乎整个语言学界都积极参与了,我们应该学习并发扬老一辈语言学家的这一优良传统,以捍卫祖国语言文字的纯洁和健康,促进中文信息处理技术的发展。

五

国家语言文字工作委员会是国务院主管国家语言文字工作的职能部门,其主要任务是按照国家的语言文字政策推广普通话、管理社会用字和中文信息处理中的汉语汉字的标准和规范,组织和推进相关的研究工作。这是一些学术性很强的工作,因此,我们同时也要加强和语言学家们的联系,要在和语言文字工作有关的行业之间(例如中文信息、计算语言学、语文教学、心理学、社会语言学、应用语言学理论等)搭桥铺路,提供一个场合、一种机会,使相关专业的朋友们汇集起来共商语言文字应用的大计。希望海内外计算机界和语言学界的朋友们多多支持国家语委的工作,我们将尽力为大家服务。

关于汉字的伟大发现

安子介

(中国人民政治协商会议副主席)

近百年来汉字有了空前的突破,大量地使用“双字”词,用了它很容易创造“叁字词”,应用于翻译外文,此其一。

中文用汉字写作,有常用“字”,用在“双字”词的第一位或第二位,意义就不同。变了位,意义又不同。另外去拉了一个新字过来,意义又完全不同,此其二。

我用统计方式在现代工商报章中找到3650个常用字,学习认识了它们,能看懂一篇文章的99.8%,此其三。

经过不断的统计研究,我发现如果认识了最常用的2000字,已足够认识文章的97.4%,此其四。

我为加速我国的扫盲工作,写了一本《安子介现代千字文——启蒙篇》。经过实践,在江西省万载县有一位六十余岁老师,教二十来人的文盲。全部学生经过二十来天学习,都已变成能识字、能写的识字人。其中有一人名叫辛桂华的女性,42岁,除耕田外,又要料理家务,因丈夫瘫痪(上海科学教育电影制片厂曾有实践记录),竟能在20日之后脱盲,此其五。

我们把这五点放在一起看,汉字是世界上独一无二的文字。我敢断言,到了21世纪,汉字必然成为世界语。这是我对汉字新的结论,胜于我以前的结论“21世纪是汉字发挥威力的时代”。

汉语书面语的分词问题
——一个有关全民的信息化问题

陈力为
(中国中文信息学会)

汉语的书面语是按句连写的,词间无间隙,因此在汉语书面语的处理中,例如统计、分析、理解等,我们首先遇到的问题是词的切分。把按句连写转换为按词连写。所以,词的正确切分是进行汉语书面语处理的必要条件,它的任何错误都将使处理结果受到或大或小的影响,有时是严重的影响。

从 80 年代初起,很多学者、专家致力于汉语书面语的自动分词[4],取得了不少可用的分词系统,但在实用的过程中,又遇到不少新问题,困扰着我们[6]。例如人名、地名、企业名、新词等未登录词[5, 6, 7]。对于这些问题,经过业界的努力,近两年来,又取得若干可喜的突破。但随着国民经济信息化的不断发展,中文信息处理的广泛地、深入地开展,对分词系统的要求将越来越高,难度越来越大。现在,汉语书面语的分词技术已经悄悄地形成了一门新兴的富有挑战性的学问。

过去经验告诉我们,我们中文信息处理技术是在不断克服困难中前进的,书面语的分词也不会例外。我们相信业界将根据客观需要,继续研究分词中的难点,推动分词技术的前进。

但是,现在我们需要冷静地想一想,汉语书面语的切分是汉语固有的属性呢?还是人们强加给它的呢?

在汉语中什么是词,到现在并无公认的定义。今天也并非讨论

什么是词的时候,但人的思维是以词为基本单位进行的。人们表达自己的思想有两种途径:语言,文字。前者叫做口语,后者叫书面语。口语中,词间有‘顿挫’(按词说出),而书面语中词间无间隙。很明显,口语忠实地表达了人们(说话人)的思想(但表情、手势等人体动作除外),而书面语则把人们思想的非常关键的信息,词间间隙,给滤掉了。因此,书面语的读者首要的任务是:使用自己的全部知识,进行词的切分,边分词边理解,把书面语滤掉的信息给补上。实际上,这对读者是十分沉重的负担,只是习惯了,误认为这是自己应该干的事。

上述书面语和口语的鲜明对照,使我们清醒地认识到,汉语书面语的词的切分问题,并非汉语所固有的,而是人们强加给它的,是人为的。若要恢复汉语原来的面貌,其办法是显而易见的:这就是由书面语(文章)的作者按词连写(词间加间隙)就是。只是所需空间增加了四分之一。在这样的书面语面前,词的切分歧义问题不见了。像‘乒乓球拍卖完了’这类的拦路虎也自动解体了(这句话指的是‘乒乓球’还是‘球拍’,难道还会难倒使用这句话的人吗?)。未登录词切分问题不见了。把一件易如反掌的事情变为一座难以攻破的堡垒,这是我们现行的汉语书面语书写规范(按句连写)造成的后果。必须引起我们的深思。

大约在 50 年代,语言学界有一次辩论:是否把按句连写改为按词连写[8],未能通过。在 1987 年中文信息处理国际会议上,本文作者也提到同样的问题[2]。最近在香山科学会议第 42 次会议上有多位学者在发言时,提到这个问题。周锡令教授在《计算机世界》上又从软件的中译本方面出发,指出这个问题的迫切性[3]。

看来,汉语书面语的书写规范已经到了必须修改的时候了。

回顾一下汉语书面语书写规范的沿革是有帮助的。在古代,汉语书面语中不要任何标点,于是标注文章成了一门高深的学问。从汉代起,读书人才注意断句(句读,‘读’音 dòu)问题。只是在大约

70 年前,‘ 五四运动 ’以后, 人们才开始使用现行的全套标点符号。可以看出, 每次改革都使原始书写者通过书面语, 传递更多的信息; 虽然书写者要多费些力气, 也增加了费用, 但由于信息含量的增多, 含糊和歧义减少了, 不仅为读者带来了好处, 社会效益也增加了。这样的大好事情当然只能留给书面语的写作者去做了。

必须指出, 汉语书面语书写规范的修改是一桩有关全民、全社会的工作和生活的大事。它的拟定和实施将遇到一系列的问题, 这些问题都要一个一个的予以解决。同时, 它也是一个复杂的系统工程, 需要有组织有计划地进行。其中最复杂的是习惯势力(例如: 看不惯、写不惯等); 它必然有形、无形地发生着制约的作用。当然, 在技术上, 也存在一些问题, 例如要分清什么是词。从时间上讲, 它不是三年五年的事情, 可能是跨世纪的大工程。但是, 只要我们有决心, 这些问题都是可以解决的。

国民经济信息化的迅速发展将迎来我国社会生活的美好前景, 并将推动信息高速公路的创建。量大惊人的信息在公路上飞驰, 为了抽取其中有用的信息资源, 人们对信息处理的速度和精度将提出极为严格的要求。面对这样严峻的挑战, 难道我们的信息处理仍然容忍被人们强加给汉语的词的切分问题继续困扰下去吗? 否! 我们还有其他更重要、更迫切的课题要去解决。

请看看英语吧。英语书面语, 除了词间有间隔外, 专用名词的首字母还要大写。书面语带来的信息超过了口语, 为信息处理提供了有利的条件。那么, 要求书面汉语恢复汉语的本来面目, 词间增加间隙, 也是理所当然的了。若是在专用名词上再增加下划线, 那就喜出望外了。但这并不稀奇, 从‘ 五四 ’前后有语体文到本世纪 50 年代, 一直就是这样的。现在, 少数古籍的整理仍然使用。

很多键盘输入系统是按词输入的, 但在完成输入任务以后, 又把分词信息抹掉了。十分可惜。

结束语: 近几年来虽然多次提到书面汉语的改革问题, 但都未

取得共识,更未见诸行动。其原因不外乎:

1. 未有充分的实践经验使我们认识到它的严重危害性;

2. 未感受到国民经济信息化的进程对信息处理的猛烈冲击。

今天不同了。我们认识到:书面汉语的改革已经刻不容缓了。而且,语言学界和信息处理界的结合也为书面汉语的改革创造了有利条件。

这样一个重大改革,必须分阶段进行。第一步,可考虑在自然科学和技术科学领域中试行,摸索经验。第二步,从小学语文教育开始,逐步推广到全社会。

参 考 文 献

[1] 陈力为. Some Key Issues in Chinese Language Information Processing and Their Prospective Developments. 1987 ICCIP Conference, Beijing

[2] 陈力为. 当前中文信息处理中的几个问题及其发展前景. 计算机世界. 1987 年. 第 21 期. 第 34 版

[3] 周锡令. 软件书籍中译本的可读性和几点建议. 计算机世界. 1995 年. 第 41 期. 第 15 版

[4] 梁南元. 再论汉语自动分词和切分知识. 1987 ICCIP Conference, Beijing

[5] 郑家恒、刘开瑛. 计算语言学研究与应用. 自动分词系统中姓氏人名处理策略探讨. 北京语言学院出版社, 1993

[6] 宋柔等. 计算语言学研究与应用. 基于语料库和规则库的人名识别法. 北京语言学院出版社, 1993

[7] 沈达阳等. 计算语言学进展与应用. 中国地名的自动辨识. 清华大学出版社, 1995

[8] 许嘉璐. 在香山科学会议第 42 次会议的发言. 1995 年 10 月 31 日

第二部分

语言学理论和汉语句法、语义研究

“名词+ 动词”词语串研究

马　真　陆俭明
(北京大学中文系)

摘要: 文章为机器判别出现在任何一个话语语流中的“名词+ 动词”词语串是不是一个合法的组合给出了条件和规则。文章首先交代了两个概念:“语法形式”和“非语法形式”。接着文章分主谓关系和偏正关系两个层面,分析说明了话语语流中的“名词+ 动词”词语串在什么情况下不是一个语法形式,不能捆绑在一起;在什么情况下是一个语法形式,可以捆绑在一起。文章分析显示,决定话语语流中一个“名词+ 动词”的词语串是不是一个语法形式,能不能捆绑在一起,主要有两个因素,一是名词或动词本身的特点,包括语音上的特点和语法功能上的特点;二是话语本身的层次构造。文章指出,前者可以通过词库(或者说电子词典)来解决,在词库中详尽描写说明每个词的语音特点和语法功能上的特点;后者可以通过规则系统来解决,文章所作的分析描写实际上给出了分析包含“名词+ 动词”词语串的话语的层次构造的规则。

§1　引　　言

本文所说的词语串是指在话语语流中一个动词跟紧挨着的一个名词所形成的词语串。这种词语串可能是一个合法的组合,也可

能不是一个合法的组合。例如:

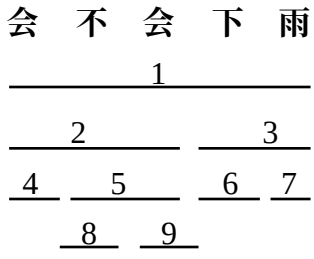
(1) 我知道小王去。

(2) 北京市的公路建设得很好。

在上面这两个句子里各有一个“名词+动词”的词语串“小王去”和“公路建设”。它们分别在例(1)和例(2)里是不是一个合法的组合?就人来说比较容易判别,稍有一点语法知识的人很快就会判断:例(1)里的“小王去”是一个合法的组合,而例(2)里的“公路建设”不是一个合法的组合。可是机器就不知怎么判别了,你要让机器来判别,就得先把怎样辨别一个“名词+动词”的词语串合法与不合法的有关规则和知识交给它。

本文就想探讨一下判别出现在任何一个话语语流中的“名词+动词”词语串合法与否的条件和规则。在正式谈论这个问题之前,有必要先交代两个概念:“语法形式”和“非语法形式”,以便先搞清楚什么是一个合法的组合,什么不是一个合法的组合。

语法形式　是指通过层次分析(即直接组成成分分析,亦称IC分析)所切分得到的大大小小的语言片段。如“会不会下雨”可作如下切分:



“会不会下雨”通过层次分析,可以得到以下九个大小不等的语言片段(包括“会不会下雨”自身在内):

(1) 会不会下雨

(2) 会不会

(3) 下雨

(4) 会

(5) 不会

(6) 下

(7) 雨

(8) 不

(9) 会

上面各语言片段是通过层次分析得到的,所以都是语法形式。

非语法形式　是指从层次分析的角度看彼此不构成直接组成成分关系的词语串。这又可分两种情况,一种是在任何情况下都不可能构成直接组合成分关系的词语串,如"我已经看了"里的"我已经"就永远是一个非语法形式;另一种是彼此有可能构成直接组成成分关系,但在有的组合里不构成直接组成成分关系,例如"会下雨"在某些话语里,如在"天会下雨"里,可以成为一个语法形式,可是在上面举到的"会不会下雨"里,"会下雨"就是一个"非语法形式",即不是一个语法形式,因为"会"和"下雨"在"会不会下雨"里并不构成直接组成成分关系(与"下雨"构成直接组合成分关系的是"会不会",而不是"会")。

§ 2　关于"名词+ 动词"词语串

2.1　在汉语中,"名词+ 动词"可能构成两种语法关系,一种是主谓关系,如"张三去";一种是偏正关系,如"进行方言调查"中的"方言调查"。什么情况下一定构成主谓关系,什么情况下一定构成偏正关系,关于这个问题,我们将另文讨论。本文是在假定这个问题已经解决的基础上来讨论"名词+ 动词"词语串在什么条件下是一个语法形式,可以把它们捆绑在一起;在什么条件下不是一个语法形式,不能把它们捆绑在一起。

2.2　在汉语中,"名词+ 动词"词语串是不是一个语法形式,