汉字的自动识别与汉字的简化

计算机科学系　黄俊杰
中　文　系　陈恩泉　张　普

当今的世界，二十世纪杰出的科技成果之一——电子计算机的科技水平、生产规模和应用程度，已经成为衡量一个国家的现代化水平的显著标志。电子计算机几乎是无孔不入、无所不能。据统计，目前世界上计算机的应用项目已超过四千多种。从某种意义上说，它正在引起一次新的工业革命。

在语言学领域，语言、文字的应用与研究工作也由于计算机的使用进入了一个新的时代——现代化时代，并且因此而产生了一些新兴的语言学科。语言文字已经不再单纯是人类所特有的最重要的交际工具，而且正在逐步成为人类与机器的交际工具，这就是通常所说的"人机对话"。过去的汉字简化工作都是从有利于人的使用着眼的，这一工作所取得的成绩是勿庸置疑的。本文想从另一角度，即有利于机器的使用出发，对汉字的简化工作试谈几点粗浅的意见。

一

无论是语言文字的应用工作还是它的研究工作，要想采用现代化手段，都必须有一个前提，即首先要把语言或者文字变成电子计算机可以识别的信息输入电子计算机，而自动识别就是文字输入的一种最理想的办法。

文字的自动识别，又叫高速输入。国外通称为"图样识别"(pattern recognition)，也就是电子计算机自己"认字读书"。它用机器来模拟人类所独具的一种处理信息的高级机能，以自动识别文字、图形及其它信息，但又远远超过了人的识别速度，真正达到了"一目十行"乃至"一目百行"、"千行"。如：美国的 IBM—1287 OCR(光学文字读取装置)和日本的 ACPET/70 OCR 都已达到每秒识别二千字符的程度。而在实验室中，据说可以达到 $5-6\times10^4$字/秒的速度。对于某些需要大量输入文字的应用工作和研究工作，尤其是对要存贮全世界每年出版的五十万种图书、十万种科技杂志和四百万篇论文的图书馆及情报部门来说，这种高速输入的手段至为必要。

文字的自动识别一般采用相关匹配法和几何特征抽取法，此外还有笔画分析法等多种。相关匹配法是把扫描所得的被识别文字的光点矩阵与预先建立的各文字类别的模板进行比较，符合或最接近哪一个字的模板即判别为该字。这种方法的特点是从文字的形状总体上作比较，它要求形体差别大，符号种类不多等条件。特征抽取法是从整个文字体系中寻找字形、笔画上的几何特征，分析归纳，并用数学上的逻辑式表达出来。这些特征叫做"一般特征"。然后通过大量模拟运算来确定每个文字所具有的这些特征的特殊标

准。这个标准是刻画每一个具体文字的总体形状的，叫做“特殊特征”。对扫描进来的每个文字信息，确定它的“一般特征”和“特殊特征”，然后用适当的数学决策方法得出判决。

但是，不论采用哪一种自动识别的方法，在识别判决前，都必须进行相连文字的“字分离”。因为当扫描器对一行文字进行扫描时，产生的是连续的信息串，必须有一个能确定每一个文字的扫描始端和终断的检测装置，将相连的文字扫描图象信息串从中切断。简单的文字分离是由输入文字之间的空格扫描来实现的。

二

能否尽快地实现汉字在计算机上的输入输出，直接关系到我国计算机应用的普及与推广，因而也就将在很大程度上关系到实现四个现代化的速度。特别是中文情报资料的检索、文献档案的存贮、现代通讯、排版印刷、企业管理和机器翻译等项工作的自动化和现代化，所受的影响更为突出，因为它们必须以汉字的自动编码和自动识别为前提。

汉字是目前世界仅存的意音文字。它的自动识别受到字数庞大、笔画繁琐、结构复杂等困难的限制，一直进展不快。而且，如上文所述，机器不但要求被识别的文字体系字数笔画要少、结构要简单，而且要求每个文字符号整体性要强、符号之间区别性要大。这就要求我们的汉字简化工作，在着眼于人易认、易读、易写、易记的同时，还要兼顾易于机器自动识别。从某种意义上说，今天要把易于机器使用摆在更加重要的地位，这就对目前汉字简化工作提出了某些新的要求。如：

（一）分批不要过勤，简化字必须相对稳定。

在“约定俗成，稳步前进”的方针指导下，汉字简化工作是分期分批进行的。第一次简化共分四批，推行简化字五百一十七个（不包括偏旁类推字）。四批之间的间隔时间分别为五个月、约两年、约一年。《第二批简化字（草案）》（以下简称《二简》）分为两表，第一表已在试用，简化一百九十三个字，类推五十五字；第二表正在征求意见，简化二百六十九字，加上类推共计六百零五字。此外，四千五百个较常用字中超过十笔的尚有一千三百个字，“希望大家研究讨论”，估且认为这是未来的第三表，或者第三次简化。

过去的改革工作，无论是形体的简化还是字数的精简，都已大大有利于人的使用和机器的自动识别。但是，我们是否也要看到分批过勤而造成的一些副作用？使用汉字的人不但要求这个工具简便，也要求它相对稳定。如果改革和相对稳定的辩证关系处理不当，已经掌握原有汉字的人就会感到不适应，动力就可能化为阻力。而且，每改动一次，不论改动的字数多少，各种教材（尤其是一套语文教材）、语文辞典、各类索引式工具书等等都要随之重新修订，否则就不适用。而一本词典的编纂往往费时多年，根本作不到随着汉字的分批简化而频繁修订。据我们所知，连国外编纂的汉外辞典，有的也因为《二简》正在试行而中断排版，要等待正式确定后再调整排版，出版发行。特别是要存入计算机的字词典，更需要文字的相对稳定。不然的话，改革一次，就要对整个汉字的编码系统和识别系统作相应调整，这就势必影响到我国的资料存贮、情报检索、机器翻译、现代通讯等一系列工作的自动化和现代化的进程。顾此失彼，这个得失需要慎重衡量。如果说在第一次简化时，汉字的自动识别等问题还未提到议事日程上来，简化工作也需要摸索经验，步子小些，分批勤些还属事出有因的话，那末，今天的简化工作是否应该

在新的情况下有新的考虑呢？

分批不要过勤，要相对稳定，这就必然会导致一次简化的数量增加。这会不会因此加重了人们的负担而不利使用呢？我们认为不会。

以《二简》第一表"不作简化偏旁的简化字"172个为例，粗略统计：象"帮、舱、蓝、龄、阎"这样的已经在《一简》简化过一次而《二简》又再次简化的字共计三十七个，占百分之二十一点五，比例可谓惊人。一字分两次简化，人们的负担实际上白白增加了百分之二十一点五，可见分批勤亦不见得一定就能减轻人们的负担。如果能够一次简化定形，则或可使人们负担相对减少百分之二十一点五，或可以增简另外三十七个字。而且其中有些简化字，在第一次简化时就在群众中使用，并非近二十年来的新发展。如蓝作兰；阎作闫；龄作令等。此外，象亍(街)、皃(貌)、卩(部)、歹(餐)、宙(赛)、仃(停)、午(舞)、亖(宣)、厷(雄)、仪(信)这样的字(亦有四、五十个之多)也几乎都是在第一次简化时就在群众中使用了。我们的中、小学语文老师二十年来不是天天在给学生改这类的"错别字"吗？它们当时之所以被当作"错别字"，就是因为"尚未正式简化"。实践证明，群众并没有因为使用这批所谓"合理不合法"的简化字而感到额外地加重了负担。相反，禁而不止，虽然"非法"，却仍流行。这不恰恰说明只要简得合理，在一次简化的数量上还有增加的余地吗？增加批量，减少批次，从而照顾到相对稳定。这一点在今后的改革中是应引以为鉴的。

周有光先生指出："不少人提出希望：《第二次汉字简化方案(草案)》经过慎重修订定案以后，汉字的笔画简化就此结束。"(见一九七八年六月十六日《光明日报》)我们认为，至少应使四千五百个较常用的汉字或常用技术字表的简化工作就此结束，这不只是人的希望，也是机器的希望，汉字自动识别工作的希望。

（二）整体性要强，区别特征要大。

汉字简化包括两方面的工作，一是简化现行汉字的字数，二是简化汉字的笔画。这两方面的工作都有利于机器识别和提高汉字输出的清晰度。因为汉字输出的清晰度与笔画的繁简程度成反比。不过，从易于机器识别的角度看，汉字的简化工作还必须注意整体性强，区别特征大这两点，否则将容易造成机器误判。

整体性强，即是要求字的结构要紧凑，横排的汉字主要是左右的结构要紧凑。如象"引"、"旧"，这样的字，整体性就不强。一旦字的两部分之间的空白大于或等于字与字之间的空白时，就会使相连字的分离造成误分。"引日"可能误分为"弓""旧"，"弓""旧"也可能误分为"引""日"。或者把"引"分为"弓""丨"，把"旧"分为"丨""日"，造成识别错误。区别特征大，即是要求形体差别明显。如象："己已巳"这样的字，区别特征就不大。若要避免误判，就要用"加权相关法"，对仅有的显示区别的部分进行"加权"，即将区别特征放大，以缩小误识率。如果我们的简化工作注意了这一点，将会进一步为自动识别创造有利条件。以《二简》的第一、第二两表为例，由于忽略了这一点，粗略统计，就造成了如下一些近形字组(简化前并不近形)：〔见下页附表〕

第一组字，形体极近，机器当然容易识错，第二组字，虽然小有区别，但因整个主体轮廓相同，一旦区别特征部分印字质量较差时，也会造成错判。在简化时，如果刻意斟酌，在不违反基本简化原则的前提下，这些近形字组有许多还是可能避免近形的。如

"囊"字可否简作"曳",这样虽多了两笔，但避免了近形。又如"幽"可否简作"凼"，用轮廓字的办法避免形声字的近形，而且与"断、继"简作"断、继"的原则也近乎一致。

第一组			第二组		
原　字	简化字	近形字	原　字	简化字	近形字
潜	汗	汗汙	芭笆耙	巴	巳
凋碉雕	刁	刀	衩杈杈汊	叉	又
嚼	啦	唯	眉媚	尸	尸
诞	诅	诅	暇	吓	吓
愚	丞	丞	事	𠂇	𠂇(高)
侵	归	白	缝	纬	纬
囊囔	市	申	幽	凼	凶
盗	资	咨	既	无	无
柬	东	东	皿	口	口(血)
道	辽	边			

* 与上列近形情况相同或类似的字组未一一列举。

（三）传统的或已简化的一些近形字组，尽管有些字笔画很少，是否也可以考虑作些改革以示区别呢？如象：

己已巳　戊戍戊　土士　未末　日曰　乌鸟　儿几　天灭　桨浆……

此外，还有一些近形的偏旁。如象：

氵讠　礻衤　才木……

这些近形字组和偏旁，不仅不利于机器识别，有的几千年来一直害人不浅。历史上这类字的改革也是不乏先例的。如："丸"本作"丸"，因与"平凡"的"凡"近形，后来改为"丸"，以扩大形体差别。又如："丈"篆文作"𠂇"，但与隶变后的"支"(𠂇)近形，所以"𠂇"变作今天的"丈"。古文"上下"本作"二二"，后来又作"丄丅"，最后才改作今天的"上下"，大约也有避免近形字的关系。

《二简》中将"没"简作"殳"。本来"没"字只有七笔，已经符合一九六〇年四月二十二日党中央关于汉字简化"尽可能使每一字不到十笔或不超过十笔"的指示精神，但《二简》仍然作了简化。这一简，不但减少了"没"字的笔画，同时也避免了与"设"字近形。又如：第二表中将二十个偏旁是"礻"的字，改为从偏旁"礻"，这就减少了偏旁"礻"、"礻"之间的误判。这些改动毫无疑问有利于机器的自动识别。

总之，在新的形势下，汉字简化工作，还要站得更高一些，想得更远一些。在不违反"约定俗成"的原则的前提下，要尽可能考虑到自动识别的特点及需要，这才会更有利于国家四个现代化的实现。希望《二简》在修订定案时，在这方面能取得新的成功。