上篇　基础理论

绪论——汉字编码如何统一

汉字正受到严重威胁，其原因来自计算机的汉字编码应用。汉字编码方案一个接着一个，对汉字基本单元的命名，取用汉字的基本单元(部件、字根……)的数量，拆分汉字的方法，差异很大。几百个方案，已经使我们拿头转向，何况现在新的方案还是接二连三地相继不断，其研制速度与过去相比已不可同日而语。五笔字型、见字识码等都花了十几年的研制时间，而现在，几个月就可以拿出一个方案来。这仅是一方面的难堪。另一方面，计算机中文应用的初期，吸收了不成熟的编码方案，占领了计算机的软硬件空间，随着计算机中文应用的逐步展开，文字约定俗成的法则就可能强迫我们接受它许多不合理的东西。

计算机中文应用的范围，正在由专业应用向普及应用发展，汉字编码方案主要是文字的计算机应用设计，它不是一种直接的商品，它需要加上软硬件技术，它需要依附，才可以作为商品。因此，不成熟的编码方案也就会附带同样的软硬件制品。如果让不成熟的编码方案全部占领现有设备的软硬件空间，要想改变就会造成巨大的经济损失，同时也会有技术上的困难。这种情况，发展下去，汉字字形结构规律将遭受肢解的危险。

汉字是中华民族几千年文化、科学的基石。现在要给它编制一套相应的代码，不应该是自己想干什么就怎么做。汉字编码方案的设计，应依据它的历史性、宏观性、客观性、系统性，考虑和研制它。任何文字的使用，都会受到社会约定俗成的规律约束，将来如果发现有问题，要想改进是很困难的。法国在制订文字之初，过分沿袭拉丁文的拼写形式，使法文吞下大量累赘的字母，扰乱读音，许多英语、法语学者，要求、呼吁改革拼写法，使文字正确表示读音，但是无法实现。文字中存在着不合理的因素，成为无理记忆的负担，强加给千百万学生，浪费人们的精力，年复一年，代复一代，受累无穷。

国家教委提出为全国中小学计算机教育选码。这是一个很好的战略决策。我们也应尽快对汉字编码的基础理论进行探索，打下基础。到现在为止，汉字编码只有研制(发明不断)，很少研究(建立基础理论)，很少提倡不同的观点，商榷、争论，明辨是非。

不进行汉字编码基础理论的研究，就无法确立选码的原则。没有选码原则的确立，选码就无据可依，无法取得统一的意见。我们经过长期的实践和探索，对多种汉字字形编码方案进行了对比研究，不仅发现了汉字字形编码方案前进轨迹，同时发现汉字字形编码方案的结构规律。笔者认为，汉字字形编码方案的设计工作，是汉字历史上的一次变革，决不应认为仅是为了键盘输入。它是我们民族的一件大事，不可等闲视之。

汉字字形编码研究是当前电脑中文应用的一个新课题。1985年以前，社会上已有不少编码方案，形成互不相让的局面，人们希望通过"评测"来解决这个问题。1986年的"评测"，没有实现人们预期的寄望。原来，"评测"也是一个新课题，是一种探索，不是一次有权威的裁决。由于"评测"的"非编码因素"太多，又被商品竞争所利用，造成了社会上一些不正常的心理因

繁，导致编码方案不断研制，花样翻新，重复劳动。几百上千的方案，大家都说自己的好，专家们无法鉴别，只能为他们签字通过。于是，谁的实力强，出得起广告费，谁的编码就最好。不成熟的文字制品，扭曲中华民族的文化基石——汉字的结构规律，如果继续扩散的话，优秀的汉字将真的背上了一个沉重的黑锅——难学难用，给代代子孙带来麻烦。

1986年评测之误是采用键盘输入速度比赛作为推荐编码方案的手段，混淆了"汉字编码"与"汉字键盘输入方法"两个不同的概念，导致产生"非编码因素"。七八年来，全国不断地组织输入比赛，以输入速度决胜负，以重码率比优劣。从专家到一般输入者，很少对此提出异议。七八年下来，"巩固"了这些"非编码因素"在人们心目中的影响。因而使一个体系有严重缺陷的字形编码方案占领了全国计算机的软硬件空间。"大旗"树起来了，不满意的人越来越多，新方案不断出现，正说明树错了旗帜。但是，却没有一个发明人知道新方案是根本无法与之竞争的。

汉字编码是一门新的科学，社会需要它。汉字编码也是一种文字工具，社会只需要一、二种，而不是几百种。正因为它是新东西，有这么多，却又缺乏理论方面的研究，我们要挑选它，就不免会出一些问题。解决问题的办法在于认识、比较、研究。认识、比较、研究是有"的"放矢，需要通过具体的编码进行，可是，我们的学术界却缺乏这种风气。自有专利法以后，汉字编码被授了专利权，变成了商品。商品竞争有一条大家都心照不宣的默契，各人都讲自己的货色好，不讲别人的货色不好。但是，汉字编码可不同于一般的商品，商品是会过时的，而编码将作为汉字的副体形式与历史同存。它是全民族的文化科学的基础工具，不能有一丝一毫的含糊和闪失。如果也把它当成不能指责的商品，就会贻害民族，贻害后代。所以，我们应该创造良好的学术环境，提倡正面交锋，开展学术争鸣。

有的人想回避对编码设计的正面交锋，想以集体攻关，研制出一个最优秀的；有的人想从各种方案中取长补短，拼出一个最优秀的。到现在为止，这两种设想都没有成功。原因有二：一、编码设计的结构体系没有弄清。集体攻关无处着手；二、汉字编码方案的结构是一个封闭的系统，设计思想形成后，其优缺点便相互依存，舍其短便失其长，无法取长补短的。

汉字编码是一门科学。科学的发展是靠规律的寻找和发现，是一步一个脚印踏出来的。汉字编码方案的研制，也是一步一步向前迈进的。

当然，许多编码设计者并没有做无为的劳动（取巧拼凑的重复劳动除外）。他们的设计，都是一次艰辛的实践，都经受了千辛万苦。但是，汉字编码是全国人民的文字工具，是全民族的信息工具，也是我们子子孙孙要继承和发展文化科学需要应用的工具。不能因某几个人付出了辛苦，就不管好坏，随便马虎地接受。不能否认，正是因为许许多多编码设计者的殚思竭虑，才踏出了这一条汉字字形编码方案的发生、发展到成熟的道路，才能够使我们建立起汉字编码的基础理论体系。

汉字字形编码对我们来说，既陌生，又熟悉。说它陌生，电报码就是一种汉字字形编码，我们已经使用了近两百年。四角号码也是一种汉字字形编码，我们小时候就知道许多人用它检索汉字。可是电脑的来临，却要求我们还要提供一种适合它使用的汉字字形编码，我们却感到有点措手不及。原来，它不同于电报或字典，电脑应用有它自己的要求：方便键盘应用，此其一；电脑中文应用的要求又不断地从低要求走向高要求；国际化、全汉字化、永久化，此其二。这些都是我们在电脑中文应用初期所没有预料到的。开初，我们还以为只要电脑能打打文章

就可以了。岂知电脑会把世界变为一个"村子",国际所有的信息交往,都因电脑的介入,变得十分方便——国际化;电脑的中文应用,将彻底改换中华民族的书写和印刷工具,节约全社会的精力。因此,要使所有的汉字都能进入,包括繁简体、日本、韩国、地方汉字和古汉字——全汉字化;电脑中文应用,还要进入中小学教育,千秋万代传递下去——永久化。这些要求,并非电脑科技一开始,我们都已经知道的,而是我们在应用与提高的过程中逐步明白的。所以,我们在这短短的十几年中,出现一些失误在所难免。

编码是文字的副体。编码应用是文字应用的另一种方式。国家希望统一,人民希望统一。我们认为,只有进行基础理论的系统性的研究,在理论的指导下才能实现统一。

第一章　汉字字形编码方案的得与失

本文论述汉字编码的思想错失，与它的成功之道，失与得是对立的统一体。希望有关专家勿以文中某些论点有犯而见责。

汉字编码是汉字历史中的一次重大转折，是汉字现代化的变革，不应认为仅仅是为了键盘输入。中文信息界需要开展学术争鸣，排除思想误区，才能真正使汉字编码走向正途。本文总结十几年来，汉字编码活动的收获和失误，但由于学浅识薄，见偏一隅，其中学术论点，纯属个人主观见解，不妥不当之处，欢迎直言批评。

一、编码方案无穷尽，汉字面临危险

周有光先生痛惜"中国丧失了一个打字机时代"。计算机的中文应用，给我们带来了一个新的希望：凭着汉字负载的信息量多，在相同的时间内，相同内容的信息量，用计算机输入中文，要比西文快。如果我们能迅速普及计算机中文应用，就可以逐步夺回失去的时间。但是，必需提供一种确实可以迅速普及的汉字计算机中文应用的方案。

几千个专家日夜奋战十几年，编制出一个又一个汉字编码方案，仍只能解决少数经过专业培训的专职输入员做计算机的中文工作。尽管在联合国科教文组织的中文输入表演中，我们的工作曾得到某种赞赏，但离实际要求仍有一个很大的距离。在联合国的表演，只是几个经过长期训练的专业打字员，离普及要求还远得很。

尽管许多编码方案的设计人确确实实是在辛辛苦苦地为计算机的中文应用忘我付出，也在一定的程度上和一定的时间内给它的普及带来好处，但又使我们陷入一个新的困境，这就是不成熟的编码占领市场。编码应用逐步铺开面拆分无规，汉字面临被肢解、被分裂的威胁。

为什么说"被肢解"？因为，汉字需要拆分才能转换为键盘符号。五笔字型的发明人王永民先生说："什么算构件(即字根)，什么不算？构件选多大合适，都因人而异，这里的根据是设计的需要。"汉字要拆为构件，而构件只能根据设计的需要，没有客观标准。即是说，汉字的构件是可以由编码设计人任意指定的，这不是很糟糕的一回事吗？如果这句话是一些没有名望的编码设计人说的，倒无关大局。五笔字型在国内占了百分之九十的市场，范围之广，影响之大是人人为之侧目的。如果几百个编码设计人都可以任意指定构件标准，几千年的中华民族的伟大创造——汉字，就将体无全肤了。

汉字的构件，即基本单元，如果历史上已经把它们搞清楚了，为它们定性定量了，汉字编码就不会这么困难。文字学是应用科学，过去没有编码的需要，我们不能苛求古人。文字学也是一门科学，也应该有可以共同遵循的客观准则。汉字编码方案的设计，是文字应用设计的延伸，也是一门科学，那么，它就必然有一定的、应该遵循的客观规律，就不应该随心所欲，"因人而异"。

字音编码如果可以使用，它的规律已经清楚，我们就不需再惹字形编码的麻烦。尽管字音编码的应用研制也已有一定的水平，许多人还是喜欢字形码。就实践经验来看，字形码的覆盖率比字音码高，作为汉字的辅助形式可行性比字音码强，应用前景比字音码好。这是因为作为汉字的辅助形式，字音有两个局限无法克服：一是语音，二是非常用字。中国人只有要求"书同文"而从没有提"语同音"的。

字形编码的困难在于汉字的拆分。许多编码设计者为了回避拆分的麻烦，采用"优选"高频部件加笔画的方法设计编码方案。例如把"我"拆为"丿、扌、单"。"戈"的横笔被抽走，余下一个不三不四的"戈"。"戈"是组字频度和使用频度极高的部件，如此任意"宰割"，谁能接受得了？

汉字的拆分，如果可以直接拆分到笔画，就不需要确定拆分的级次。正因为它不能直接拆到笔画，我们才需要先明确汉字的层次。

汉字可以分为三个层次：汉字——部件——笔画。笔画组成部件，部件组成汉字。笔画是最小的单元。汉字的层次最高，也有单个笔画参与，如："一、乙"。部件界于汉字和笔画之间，有一部分是有字义的汉字，也有一部分是可以独立拼字的笔画。这说明汉字的层次是不能"一刀切"的。

汉字的使用使全国十几亿不同民族、不同语音的人们团结为一个整体，却因计算机的中文应用而采用不同的编码。如果时间一久，计算机普及到家家户户，键盘真正取代了人们手中的笔，各种编码各有自己的应用范围，拥有自己的小"山头"，"书同文"就会被彻底分裂。全国建立计算机的统一网络，必将会有较多的麻烦。如果那时要强行实施统一，就会有比较大的损失。

二、方案不断增加，好坏不能识别

十几年来，我们研制出几百个汉字字形编码方案。一次次的实践，使我们对字形编码设计方法的掌握，和汉字字形结构规律的认识不断提高，使计算机的中文应用顺利开展。短短的十几年时间，全国人民对计算机中文应用的迅速普及，对键盘将最终取代我们手中的笔杆已深信不疑。

由于我们已经处在一个计算机中文应用全面普及的前夕，而而对的却是一个众多编码无法选择的局面，无法确定谁优谁劣。报纸、电视，经常出现发明汉字编码的报导。有的人确确实实是经过艰苦研究，为字形编码的前进铺砖垫石；有的码却是东拉西凑取巧而成。鱼龙混杂，难以鉴识。人们对于早期的编码发明，采访报导头版头条，大篇幅文章捧上了三十三天。后期的发明人却需要自掏腰包，请客、送礼、拉关系挤进某个角落头。同样是汉字编码方案，个别人已腰缠万贯，许多人却破家荡产。尽管这样，编码方案还是出个不停！

支秉彝、王永民等先生都是以十几年的辛勤研究才搞出一个方案，现在有的方案却只要几个月就可以研制出来。人们似乎对编码的研制方法已了如指掌。所以，这几年来，一下子爆出几百个，而且还是陆续不断地在"爆"，其势头更甚于前。

编码方案的不断增加，所有的设计者都想获得国家管理部门的支持，这就难为了信息界的专家。这里要会而，那里要鉴定，开会签字，忙个不停。只顾应付编码发明，却无法区分优劣。

拿不出评判的标准，如何市定它们的优劣？然而，中国人又不需要这么多的编码方案！

作为编码的发明人，自然觉得自己的方案最好，谁也不会花几年时间，去研制比别人差的编码！何况几年的辛苦下来，发明人对自己的方案中每个字的代码，已滚瓜烂熟，就会觉得自己的方案，易学易记，对别人的方案，就会横竖不顺眼。把编码方案做成软件，鉴定，宣传，推广应用，不仅要花更多的精力，还要陪上血本。所以，编码发明人自己是无法作出抉择的。许多人寄希望于信息专家。信息界却认为是"各有千秋"，有的人则想从中"取长补短"，拼出一个最好的来。

为什么这么多的编码方案无法辨识？因为，汉字编码的基础理论尚未形成。汉字编码的性质，编制编码的目的、方法、要求，都未弄清楚。

三、编码乃是变革，不是专为输入

一般来说，大家都认为汉字编码就是为了键盘输入，都从键盘输入的角度考虑对编码的要求，其结果是越来越乱。这说明我们把它的内容弄错了。

汉字编码是"按照一定的规则，给汉字[汉语词语]集内的元素，编制相应的代码。"其工作内容有三：

一、制订可"按照"的、"一定"的"规则"；

二、确定"汉字[汉语词语]集内的"元素"；

三、想出为这些"元素"编制代码的方法。

拼音字母是汉语的元素，已经确定，如果音码可行，人们就不会再设计字形编码。字形编码的困难是明摆着的。几千年的汉字史中，没有现成的可"按照"的"规则"，因为，汉字编码是汉字的现代应用，我们不可能在古代就预先为它制订"规则"。"规则"未曾确立，"元素"自然无法确定，被编制代码的"对象"只能先主观拟定(即"优选")。

汉字史上没有汉字的字形"元素"，现在我们要通过字形编码来确定汉字的字形"元素"，这无疑是汉字应用史上的一次大转折。因此，可以认为汉字字形编码方案的设计，是要把汉字作一次人为的改造，使它适合现代科学、文化、教育事业发展的需要，而不仅仅是作为一种输入法，应该认为这是汉字发展过程中的一次变革。所以，陈爱文先生说："汉字编码是汉字的的辅助形式，是中国文字的第二种形式。"

汉字编码既然可以认为是汉字的辅助形式，那么：

(一)汉字编码必须能够代表全部汉字；

(二)汉字编码的基本单元必须能拼出全部汉字；

(三)汉字编码的从属关系主要是文字学的。

(四)要从汉字自身发展过程中寻找拆分规律，只有这样，才能适用于全部汉字。

因此，汉字编码的应用是宏观的；要适用于所有使用汉字的领域，要适合所有使用汉字的人们。同时要与识字教育相结合，还要能继承所有的汉字文化。

五千年来，随着民族文明的进展，汉字从古代的几个简单的象形符号(自隶书后演变为笔画结构符号)不断地拼合，增多至五六万个，记载了中华民族五千年的文化。现在要使它适合现代应用，需要把它们拆分为部件。从拼合到拆分，是一场艰难的变革，必须要求这种变革，符

合汉字自身规律。

寻找汉字规律，许多人从事汉字频度和字根频度的统计。

四、编码设计困难，频度不能解决

上篇所述，归纳为一：汉字编码的实质，是要为汉字设计出它的字形元素，并使用编制代码的方法使之成为一套有序的符号系统，产生这套符号系统的方法是要将汉字拆分为部件，将部件加以分类，采用有序的键盘符号为代表，使汉字序性化。汉字拆分为部件后，笔画数少了，就有利于识字教育和书写规范；汉字序性化了，排序和键盘输入都一起解决了。所以，决不能认为，编码方案的设计，仅仅是为了汉字的计算机中文输入，而是汉字现代化系统工程的一个方面，单一地考虑某一点，就不易成功，也容易出偏差。在这个问题上，一百多年来我们已经有很多经验教训，如汉字拼音化、汉字拉丁化、速成识字等。

在计算机技术进入的初期，我们还摸不清如何设计编码方案，如何更好地使汉字在键盘上打出来。为了早日使它在计算机上使用，研究汉字和部件的使用频度，作为一种应急措施，是很有必要的。但是要把它当作一种最科学的、最根本的方法，作为一种字形编码的设计或优劣分析的依据，那就错了。要建成一个整体的汉字符号系统的设计方案，不仅仅靠几个常用字或高频字根。这个符号系统，应该能够反映全部汉字信息，能够使计算机继承中国全部汉字文化遗产。这才是汉字字形编码方案设计的思想基础。汉字字形编码不应只是常用字（或高频部件）的编码，应该是全部汉字的编码，即能包含所有汉字（我们提出这个条件的意思是指汉字的拆分方法和分类方法，都应该具有可扩充性。）因为，从书写工具到印刷工具，都将为一个单一的键盘所取代，难道我们只需要一部分常用汉字或一部分高频字根，建成汉字的符号系统就可以了？汉字编码，是将汉字集的元素编为代码的设计，应该寻找它的设计规律，而不是从它的元素中选取使用频度高的编制代码。

把字根作为汉字的基本单元进行频度统计之误的第二个问题是：字根的标准是什么？没有大小的标准，就不知道多少！就是说字根需要定性定量，才能进行统计，否则，这个统计数字就不准确。

我们认为，汉字的使用频度统计可以为输入软件编制和键盘安排提供基本材料（如简码的安排），但却不应该是汉字编码方案设计思想的基础。现在，软件功能可以使一级简码（即一键一空格）打出一百多个高频字，占常用频度的百分之五十以上。就是说，在输入时，有百分之五十是按形或音，或其他方法安排的，不是按编码方法打出来的。例如表形码的“我”，安排在“W”键，一键一字。如按表形码的编码方法，应该是“J.F.S”三键，才能打出来。所以，部件频度统计对一级简码来说，没有什么作用。

五、部件（字根）没有标准，频度实难作准

部件没有选取的标准，各人都有自己的“队伍”，“队员”多少可以主观决定。所以频度也就很难作准。使用无法作准的数据，作为编码方案基本单元采用的依据，设计方法就不科学。

毛笔发明之后，篆书全部改为隶书，不是把几个常用篆字改为隶书，而留着几个不常用的篆字不予改造。五千年的汉字文化，是靠所有的汉字才能反映和继承的，我们不能因它们的频度高低而任意取舍。因此，从汉字（或部件）的使用（或组字）频度来考虑设计汉字字形编码，在思想上是不够成熟的。不成熟的设计思想就会产生不成熟的字形编码。

既然汉字编码的基本单元是部件，就应首抓部件大小的标准而不是使用频度。我们所说的部件大小的标准，其含意并不是说要定几个以下的笔画才能算是部件，而是指部件需要有准确的定义。我们就可以遵照这个定义来制订拆分的原则和规则。运用拆分规则拆分汉字，就可以得到汉字部件的总数，这就是部件的定量。汉字集的元素是汉字形码的根本，根本不立，基础不固。

六、混淆两个概念，评测走进误区

1986年举办了一次评测活动，其主要指标是输入速度和重码率。这次活动在一定程度上，推动了计算机中文应用的普及和促进了计算机软件水平的提高，也同时把人们对汉字编码的认识带进了误区。这个误区就是后来的以输入速度比高低和以意码率比优劣的设计思想。然而，经过多次的速度较量，和许多无重码编码方案的出现，我们才明白输入速度和重码率不可能解决汉字编码的优劣测定。

回顾这次评测活动失误的原因是用汉字键盘输入方法来评测汉字编码，混淆了两个不同的概念所致。汉字编码是指“按照一定的规则，对汉字[汉语词语]集内的元素编制相应的代码”，而汉字键盘输入方法是指“运用某种编码方案、键盘设备及计算机资源，由操作者向计算机输入汉字的方法”。“键盘设备、计算机资源、操作者”都是“非编码因素”，用它们来决定汉字编码方案的优劣，推荐编码方案，“评测”不能不走进误区。

所有编码方案设计者都指望汉字编码专家委员会能够就此分出上中下来。十几年过去了，专家们却认为是“各有千秋”，无法选择。

编码方案的设计，由于缺乏基础理论的研究，弯弯曲曲地前进是不可避免的。张普教授在1986年的评测之后，就发表文章指出评测中存在着“非编码因素”。同时指出“首要的工作是对汉字的部件和结构进行基础理论研究。”但时至今天，速度和重码率等“非编码因素”仍为一些人奉为测定编码方案优劣的标尺。

许多编码方案除研制输入软件之外，都想编制字典，使编码与识字教育紧密结合。汉字编码，主体是汉字，客体是编码。用编码编写字典，也说明了这种主客关系。编码工作应该是维护主体的规律，作兼容客体的设计，不是破坏主体适应客体。有的人却认为编码方案只是为输入而用的，要汉字“服从”信息，反宾为主。思想视野的局限，使许多人对汉字编码内涵和规律的认识水平得不到提高，导致编码方案无法鉴识而数量却不断地增加。

七、总结得失，继往开来

（一）字形编码的研制和输入软件设计的努力，使计算机应用技术及时得到普及和提高。

(二)进入市场的字形编码没有成熟,市场使用会使这些编码具有相对的稳固性,造成对汉字内在规律的破坏,使汉字遭受肢解的威胁。

(三)字形编码的不断研制,能够使它从实践升华为理论研究。

(四)字形编码的不断研制和社会应用,由于无法鉴识,造成了汉字使用的混乱。

83年中信会汉字编码专业委员会编印的汉字编码参考资料收集论文27篇,大部分属于编码方案,其中附带一些设计方案的研究。唯一一篇不提方案专攻基础理论的是张普的《汉字部件分析的方法和理论》。另一篇原益中《如何提高汉字输入速度》,虽然没有提出编码方案,但属于编码键盘应用。1987年中信会的中文信息国际学术会议的论文集,只有一篇陈爱文的《如何设计一套全汉字、全用途的编码》的基础理论专著。钱伟长的宏观编码有一部分是很有价值的基础理论,可以认为,这个方案,为部件编码的突破创造了条件。1991年中信会的学术论文70篇,大多数是论键盘输入方法的,只有一篇王力德的《对汉字字形规律的再认识》题目是属于编码基础理论研究的,实质仍是键盘输入。另两篇提出了两个编码方案。一篇是武震声的《汉字输入的瓶颈是汉字的拆分》,以探讨拆分为契机,提出一个设计方案;另一篇是李挺进的《汉字拼形方案》。我们对这两方案的认识是:有进展无突破,放在以后再作评述。由是可知,1991年的中信会的学术论文报告会,汉字编码方案设计的方向,是偏重了键盘输入而忽略了编码本身的基础理论研究。

八、难题在于拆分,回避不是办法

计算机的来临,信息时代的大门能不能为中国人敞开通行无阻?就要看汉字能不能构成一套有序的符号系统,方便排序检索和键盘应用,同时也不能背离汉字的自身规律,影响识字教育。

要使汉字能够很方便地在键盘上使用,必须拆分为部件。"拆分"不同于拼音化的"改造","拆分"是汉字发展历程中的一次转折。汉字从古代简单的符号,拼形拼了几千年,拼成了五六万个,记载下浩瀚的汉字文化信息。计算机技术使文字应用的步伐急剧加速,使用量几倍地递增,以前缓慢的整字手写应付使用,已不适合当前的信息爆炸,汉字需要的拆分为部件后,在键盘上应用。

上万个汉字拆分为几百个部件,对提高识字速度、统一书写规范带来极大的好处。故清代文字学家王筠说:"苟能分一字为数字,则点画必不可缺,易学而难忘矣。"笔画少了就易学易记易规范,可以进一步发挥拼形优势,提高儿童的组合和形象思维能力。

许多编码设计者不敢正面相对,采用"优选"代替,使客观的拆分,陷入了主观的樊篱。"优选"的依据是部件使用频度统计。

有的人也看到了拆分是个难题,提出了另一种设想。如武震声的《汉字编码的瓶颈是汉字的拆分》,"这一问题一旦解决了,则汉字输入问题便迎刃而解……"。然而该文并没有提出"符合汉字固有规律的汉字拆分法",而是想依靠"规范偏旁"、"最简独立汉字"、"规范笔画"实现拆分。国家没有"规范偏旁"、"最简独立汉字"、"规范笔画"的规定数量,其结果还是由设计者自己"优选"决定。这是一对连环圈,从这个"圈"跳出去,却落进另一个"圈"。现在有的设计者也曾这么说过,要国家先制订一些标准部件,然后由他们去编制代码。如果国家能够解决标准部

件的问题，难道就不能解决编代码的问题？

由于汉字字形编码的不断研制，新的编码方案不断"诞生"。市场上的宣传与竞争，更增添了选择编码的复杂性。选择的困难与汉字拆分的不规范，就会造成许多不良的后果：

(一)背离汉字字形的统一使用；

(二)损害汉字字形的规范化；

(三)影响办公自动化的进程；

(四)浪费计算机的空间资源。

因此，必须要求编码方案的设计者拿出一张汉部件清单，条件是：

(一)可方便地应用的数量；

(二)只需要很少的记忆量；

(三)能建立客观的分类系统(即科学的集合方法)；

(四)适合计算机快速输入；

(五)适合中小学识字教育，不背离汉字书写规范。

但是这一切并不很容易。新科学要依仗新概念的产生。

九、提出科学新概念，汉字编码大跨越

在笔画结构块的概念提出之前，好多人为了区别重码，采用字型信息，研究字型结构，根据汉字各个"块"的排列方位，把汉字分为17种、49种、88种、100种图形，称为"型"。这种分型，由于方法不同，或者由于汉字参与数量不等，类别差异太大，没有应用价值。有的人就干脆把它分为三类：上下型、左右型、其他型，采用1、2、3作为它们的代号，作为某些汉字字形编码的区别信息。分型采用了与汉字字义完全无关的方位关系，即"块"与"块"之间的关系，使汉字从字义的纠缠中解脱出来。

钱伟长先生率先将这些"块"，命名为"笔画结构块"，并提出"互相分隔、相对独立"的划分标准。"笔画结构块"是文字学中一个全新的概念。它不同于部首和独体字，不受字义的束缚。"块"是聚在一起的笔画结构关系。"块"的概念的提出，是汉字字形编码一个新历程的起点。有的人由于长期圈困在字义的框子里，舍不得丢这个老"宝贝"，对某些没有字义的"块"，总觉得别扭，这并不奇怪。因为，笔画结构块是一个新事物，旧观念与新事物总是有格格不入的感觉的。

笔画结构块是应计算机中文应用的普及而提出的，它的好处有：

(一)"块"从属于笔画结构。笔画结构有很明显的规律性。

(二)"块"有很明显的内聚性，相对的独立性，可以依此探索出拆分规则。

(三)有了拆分规则，便可以拆分全部汉字而得到部件清单。

计算机中文应用，引来了几百个字形编码方案的设计，踏出了汉字拆分之步。汉字字型分析导出'笔画结构块'，摆脱了字义束缚。

陈爱文先生对笔画结构块作了进一步的分析："一、部件是一个'笔画结构块'，那么它应该有两个以上的笔画，而且这几个笔画应该形成'块状'。二、部件与部件之间是彼此分隔的，那么相交叉的笔画群当然是一个'块'，一般不能分入两个部件；相粘连的笔画群一般也应该是一

个'块'，一般不能分人两个部件。三、分隔开的两个'块'不能称作一个部件。"

"我们想把上面的意思弄得更具体一点。有的笔画和笔画之间，虽然是彼此分隔的，但是应该算一个部件，因为它是'块状'的。例如'氵'、'冫'、……，这些笔画群中，存在一种内聚的形式使人觉得它们是一个'块'。又如：'犭'、'灬'、'𠂆'、'二'、'三'……等，这些笔画群中，存在一种平行的形式，也使人觉得它们是一个'块'。""部件是汉字中由分隔沟隔开的笔画结构块。笔画结构块有如下几种类型：一、相交叉的笔画是一个结构块。相粘连的笔画，一般属于一个结构块。三、布局匀称的相分离的笔画是一个结构块。四、封闭框内部的笔画，如果跟外框是粘连的，则合起来算一个结构块。框内的点笔一般附属于外框。五、单独的点笔一般附属于它附近的结构块。"笔画结构分型，使汉字部件组成一个整体的分类体系。在此基础上，陈爱文先生设计了汉字表形符号编码(简称表形码)，使汉字编码的设计，作了一次大踏步的跨越。表形码的推广应用已有七年时间，尽管编码很好，所知者寥寥无几。其原因十分复杂。笔者以后再撰文细谈。表形码是否已经完善，完全可以作为汉字的元素，提供给全国人民作为识字教育的工具了？笔者认为：否也！表形码可以作为研究的基础，陈爱文先生虽然已经建立了一部分基础理论，但并不完整和完善。因此，仍有许多地方值得商榷和改进。其理由笔者将撰文详谈，以后陆续刊出，请读者注意。

十、部件分类序性化，排序输入同步上

现在，使我们感到的是信息量对社会进步的决定作用。文字信息应用的多寡，成为社会进展的一种标志。使用拼音文字的国家，早就普及打字机了，面我们的打字机，只能给专业人员使用。周有光先生才惊呼："中国丧失了一个打字机时代！"

计算机的中文应用为什么要编码？因为计算机只有很少数的几个键符，无法直接使用上万个汉字。没有字母的汉字，给现代文化普及、社会管理和科技应用带来很大的困难。例如每个图书馆、档案馆的管理，都得自己编制一套检索系统，专人管理，难以统一。

计算机技术使时代的步伐加快，摆在我们面前的任务是全国要实现计算机网络管理；是"计算机要从娃娃抓起"(进入中小学教育领域)。以前人们觉得可以凑合着应付的汉字笔画部首排序，已经跟不上时代的要求。拼音排序更无法实现这个宏伟的目标。唯一的方法是寻找汉字字形的"元素"，研制一套有序的汉字拼形符号系统。

钱伟长先生的宏观码首创宏观集合法，把形象类似的部件归为如"冂、冃、冂、⺆、匚"归为一类，又提出了"笔画结构块"的新概念。陈爱文先生则总其大成，设计《表形符号编码(简称表形码)》，利用31个有序的键符与部件类的特征，构成百分之九十以上的形象对应(映射)，基本上完成了汉字部件的序性化，汉字终于走进了有序性的行列。

汉字有了序性，全国的统一排序检索和统一的计算机网络管理就通行无阻。所以，只要汉字有序性，能排序，就能输入。

十一、建立基础理论研究，完善汉字字形编码系统

钱伟长确立"笔画结构块"的概念，陈爱文的《汉字编码的理论与实践》，《汉字字形学和表形符号编码》系统地建立了汉字编码的基础理论，才有了我们现在的"草案"。（其中许多编码实践家如朱邦复、李金铠、支秉彝、郑易里、王永民、陈代字等都是功不可没的。）

汉字字形编码方案的设计是一门科学。科学的诞生，都有一个从不成熟到成熟的过程，都是一个不断地实践，总结，提高的过程。在这短短的十几年来，为了应付实用的需要，为了计算机中文应用的普及，为了汉字字形编码的成熟，在社会上流行的一些不成熟的字形编码，确确实实做出了卓绝的贡献。现在，计算机中文应用正处在由专业应用转向普及应用的前夕，我们必须及时把握这个时机，做好汉字字形编码基础理论的研究：

（一）汉字发展的历史规律。汉字从甲骨文、金文、大篆、小篆、隶书、楷书、简化字，几千年不断的演变，原因？现在，我们要使它分成较小的单元，方法？

（二）汉字字形编码发展的规律。汉字字形编码发展中各种编码的相互关系、发展方向和结果。

（三）汉字字形编码的结构设计规律。始点、相互关系、结点的层次剖析和系统分析。

（四）汉字编码的应用范围、相互关系、社会效益、前景等分析。

总而言之，根据汉字编码方案的性质，它绝大部分应从属于文字学的范畴，应以文字学的规律为准绳，以信息应用的要求为目标，切不可主次颠倒。

十二、悟以往之不谏，知来者之可追

忽略汉字编码方案基础理论研究的原因是陋生；另一个原因是汉字键盘输入技术的进展较快，使有的人认为可以用软件技术来弥补汉字编码方案的不足。但是，加强楼房的牢固度是不能防止沉陷的，根本是加固基础结构。

汉字编码是汉字键盘输入的基础，"加固"这个基础的方法只有三条：

（一）确定汉字的形素（音素已经定性定量）；

（二）设计一组完整的规则；

（三）研究对比编制代码的方法。

上述没有一点与键盘输入有关，现在有的人要用"键盘输入"套着汉字编码走，好象非有键盘输入，否则就成不了编码。他们没有理解：只要汉字有序性，能直接排序检索，就可以用于键盘输入。而汉字键盘输入方法（或系统）却非得有汉字编码不行！因为，它一定要"运用某种编码方案"，不然，就成不了"汉字键盘输入方法。"

作为汉字编码方案要研究的三点，笔者试谈以下看法：

（一）形素命名为部件（或拼形字母）。命名后首先得给部件定义。部件的定义要根据汉字应用的历史性、宏观性、客观性、系统性为内涵，探索出汉字的拆分原则和拆分规则。

（二）根据拆分规则确定汉字拼形字母的定量，即得出拼形字母清单，根据清单进行分类。

(三)按类给出代表字母,按代表字母的形或音安排键符对应(映射)关系。

我们如果把以上条件加以综合,可以得出以下的结构图式:

以定义确立拆分原则和规则;

以拆分确定部件(形素)的总量;

以总量建立分类系统;

以分类编制代码。

以上工作次序不能颠倒!这就是"一组完整规则"!

这是笔者研究了多种字形编码方案后,以及仔细推敲国家标准局关于汉字编码方案和汉字键盘输入方法的定义,结合键盘输入的实践后,得出的结论。

过去的已经过去,有得有失,失即是得。现在正是计算机开始普及家庭的起始时机,机不可失。抓紧时机,加强汉字编码基础理论的系统性的研究,为民族、为子孙的书(键盘中文输入)同文,共同努力,为时未晚。

第二章　汉字字形编码发展的轨迹

【提要】本文论述了几个比较著名的字形编码方案的发展轨迹，阐明汉字字形编码方案的实践进展，研制汉字字形编码方案的困难被逐步认识和发现，导致编码设计的最后突破。

引言——谈选码

邓小平同志说："计算机教育要从娃娃抓起。"要从娃娃开始入学的时候，就抓计算机教育，首先就是中文输入教育。作为中文输入教育的基础，就是汉字编码。汉字编码分为两大类：字音码和字形码。本文着重研究字形码。

国家教委决定为中小学计算机教育选码，是一个非常及时的战略决策。但是，应该如何选码，各人想法不一，其原因在于有的人没有看到汉字字形编码的每一次实践，都是逐步向前迈进的。他们认为，各种方案，各有千秋，难分优劣。如果真是这样，选码就要成为空谈。

现在对选码有四种意见：

一、选择各种编码方案的优点，拼出一个最优秀的来；

二、集中专家攻关，设计出一个最优秀的；

三、把社会上应用面最广的选出来；

四、寻找汉字编码方案的发展规律，找到一个最成熟的。

本文是属于第四种的。本文认为：汉字编码方案的设计已经从发生、发展趋于成熟。汉字字形编码方案设计的规律已被发现。

汉字字形编码方案的设计是一件非常重要的工作，也是一起具有深远历史意义的大事。正因为它是一桩千秋大业，所以，许多研制者心甘情愿地为此付出他们无以数计的日日夜夜。

计算机将很快普及，键盘打字将取代我们用笔写字的方法，因此可以认为汉字编码是汉字的辅助形式，是汉字的第二体形式。以后我们使用汉字编码，就象使用汉字一样，所以，汉字编码方案的设计应该是很严谨的。

我们是一个统一国家，用的是一种统一的文字，编码是文字的副体，是文字的另一种形式，如果各人使用不同的编码，会使统一遭受损害。编码越不成熟，造成的损害越严重。不成熟的字形编码拆分汉字，不依据汉字结构规律，主观乱拆，人为地破坏汉字使用的严谨性，对汉字来说，是一种"肢解"的威胁。这就是"编码污染"。

汉字字形编码方案的设计工作，是一种文字的设计工作。我们是为方便学习而设计它的。键盘中文输入是它的应用。曾有一段时间，我们把注意力集中在它的键盘应用上，而忽略对它本质的揭示。用输入速度或重码率来决定编码方案的优劣，致使有的人以为汉字编码的主要问题是速度问题，好坏可以用打字员的熟练和键盘安排来解决。把编码方案的应用功能

当作本质，造成当前编码方案不断增多而好坏不能识别的困顿状态。这种情况如果不能在近期解决，各种编码方案的“战国割据”局面得到延续，将对全面普及计算机中文应用，和实现全国计算机网络管理，会产生不可估量的消极影响。

现在，计算机的中文应用，已处在由专业打字转向家庭普及的时机，抓紧这个时机，做好中小学生的选码工作，为摆脱编码方案的“战国割据”，防止“编码污染”，统一键盘的汉字应用奠定基础。

汉字字形编码的键盘应用要尽快统一，其前提是思想认识的统一，即对汉字字形编码性质的认识要统一。上面已经提到，汉字编码是汉字的辅助形式，是汉字的第二种形式。如果这个提法能够得到共识，各种不同的编码方案的统一就比较容易。

本文认为汉字字形编码方案的研究，应从历史发展的角度来观察它和研究它，才能摆脱“各有千秋”的优劣不可知论。

汉字编码的研究和研制，应该是一门科学。每一个方案的研制，仅仅是一次实践，一次探索。它的基础理论未有形成。没有理论指导的实践，是一种盲目的实践。没有理论的指导，也无法区别每一次实践的得失。

几年来，我们着重研究了几种比较有名的方案，发现了汉字字形编码方案的设计在不断前进，并且已经从发生、发展走向成熟。

各种不同方案的核心问题是汉字的基本单元问题。汉字字形编码的统一，首先是汉字基本单元的统一。汉字的基本单元就是汉字的部件。如果人人都可以自由地“优选”部件，那么就永远无法统一。所以，本文认为要想统一，就必须要求任何一个方案在设计时，首先要对该方案中使用的部件进行定性定量。

什么叫做部件的定性与定量，如何才算定性定量？我们在下面已经提出答案。

本文把汉字字形编码的发展分为五个阶段。困难在于部件(字根)编码阶段。这个阶段的方案多而复杂，无法一个个进行研究。我们经过分析，认为可以分为三类：笔画部件码类、形音码类、纯部件码类。这样，复杂的部件码方案，就可以排队归类了。

本文试从历史发展的角度，俯观整个汉字编码方案逐步成熟的过程，探刻剖析了几个比较著名的汉字编码方案，阐明个中道理，并不是对某种编码进行评判，完全是为了实现国家教委“八五”攻关选码课题，希望有关设计人谅解。

由于汉字编码是一门社会迫切需要而我们却比较生疏的新学科，每一个设计人都象“瞎子摸象”，每个人都摸到了一部分，虽没有窥其全貌，但却为勾勒全局作出了贡献。

本文认为：汉字编码方案的设计，从电报码到表形码，已经趋于成熟。成熟的果实需要采摘，否则就会“腐烂”，造成不应有的“浪费”。汉字编码是全国人民的历史财富，如果任其“浪费”，我们是无法向子子孙孙交代的。

一、从电报码到四角号码

电报码用四个数字代表一个汉字，这种方法无道理可讲，学习的人要死记硬背，叫做无理编码。它启发了我们使用数目字可以增强汉字序性的思想。二十年代，王云五先生发明了四角号码。他截取汉字四角笔形，给以数字代号，用来排序检索，果然非常方便，比之部首笔画检

字，速度要快得多了。由于汉字四角笔形过于简单，重码太多，不能作为电脑编码使用。但它启发了我们使用数字和笔形给汉字编码的思想。当电脑技术传到我国，就产生了笔形码。

二、笔形码阶段

使用四角号码是不能对汉字进行拆分的，编码的局限性很大。笔形码把汉字四个外角的笔形数字代码，转换为汉字自身的笔画数字代码，这是由外向内的一次阶段性的转移。汉字可以实现拆分，编码的局限得到开拓。

用笔形编码的方案很多，如徐漫的“汉字六角编码方案”，徐浩平的“汉字基本形母编码法”，申敬文的“汉字五码检字法”，姚鸿滨的“鸿钟码”，李金铠的“笔形码”等，其中以李金铠的“笔形码”最为出名。

把汉字拆分为笔画，代之以数字，是一种简单明了的编码方法。汉字的基本笔画不多，易学易记，与四角号码相比，它向着整体性迈进了一大步。李金铠先生的笔形码，是一种影响较大，有一定使用面的编码，我们可以以它为代表加以讨论：

笔形码把汉字分解为笔画，笔画分为六类，再加上二个部件，十、口（严格地说应称为结构）共八类，列表于下：

一	丨	丿	丶	乚	𠃌	十	口
1	2	3	4	5	6	7	8

笔画类少，代号少，基础记忆单位少，学习的入门阶段很容易。但是，汉字有几万个，基础代号过份地少，必然造成取码规则繁杂，规则繁杂又会使学习困难；

1、取码规则不按笔顺，按笔画的高低分先后，它对消除重码有很大的好处，但它是不规范的。我们书写的顺序，已经几千年了，现在要从高低来看顺序，思想上不易接受的。在教学上也造成一定的困难。

2、有的汉字笔画高低不明显，如：“我”最高笔是左边的撇，还是右边的右钩？“正”，是左边的短竖高？还是右边的短横高？

3、笔形码除笔画外还包含两个部件（结构）“口”“十”，交叉作为一个拆字单位，也造成拆字中的许多繁杂情况。如：

羊　十　丨　一　十　口　　曲　十　丨　口

耳　一　丨　十　一　一　　舟　十　十　𠃌

这些字，都要逐个学过才能知道它的拆法。

4、有的笔画在不同的字中会发生变化，如“月”的第一笔是撇，代号为3，在“肖”中它的第一笔是竖，代号为2。又如，“雨”的左边的竖笔，代号为2，在“露”中，雨的这个笔画变成点笔。诸如此类，都要逐个学过。

5、有的汉字笔画多达二、三十笔，当然不可能逐笔取码，这就迫使笔形码规定许多省略取码的规则。规则如下：

⑴超过六笔的字最多取六码。合体字一分为二，各取三码。如“需”拆为：

雨　1　2　4（左边一竖作点）

而　1　3　2

这里就出现新的困难了——如何把汉字一分为二？

⑵对上下分解不明确的字，优先以分成两个字为准。如：

意：音　心

坐：土　从

象"坐"拆为"土"、"从"，这是要专门学过才能用的。还有如："章"字，既可拆为"立""早"又可拆"音""十"，这又要一条规则来解决。

⑶不能分成两个字的，先以分成一个字，其次以分出一个部首为界。如：

爱：爫 —— 友　　唐：广 —— 言

皋：⺍ —— 夲　　率：⺮ —— 十

执行这条规则，又要记住一张部首的清单。对"衰"字，先拆出"亠"，而"率"字，却不拆出，这损害了规则的一贯性。而"夏"字，则更难了，既拆不出一个字来，又拆不出一个部首，里头有一个"目"字，是藏在中部的，如果拆出来，就要一分为三了。

⑷对于左右三分的字，同样先以分出一个字来，其次再以分出一部首为界。如：

树：木—对　　衡：彳一？

⑸第二部分仍可一分为二的按每部分取三码的规定，分部顺延取码。如：

像　亻　𠂒　豕　　倘　亻　冂　口

　　32　314　82　　32　243　8

薛　艹　鱼　羊　　倏　亻　𠂒　已

　　72　358　43　　32　343　56

溢　氵　𠂒　皿　　蔷　艹　土　回

　　46　431　26　　72　743　88

规则复杂到这一步，笔形码已经不是笔画编码，而成了先部件再笔画的二次拆分了。总的来说，在入门阶段，笔形码是很容易的。但把整个编码学会，并不是几十分钟可以学会的，有许多字都要逐个记忆。

笔形码只用了八个键，要把成万个汉字区分开来，每个字要用六码，输入速度就不能快，人有双手十指，正好适应键盘上的二、三十个键，笔形码只发挥了八个键的作用，等于只发挥了一只手的作用，就无法达到应有的速度。

笔形码作为进入电脑中文应用的第一道阶梯，开始拆分汉字和使用汉字部件，虽然不是十分完善，但它的历史功绩是不可泯没的。

张普教授指出："尽管每个汉字确实是由点、横、竖、撇、捺、折等若干笔画组成的，但是我们认为拆到部件一级的部件拆字法，更符合汉字的造字法和造字历史。

"古人称独体为文，合体为字。最初的独体的'文'，是象形的。它的结构和笔画并没有一定之规，而是随物质本身的曲线来构象，与图画相近，直到隶书出现，笔画才大大改观，真正形成平直方正，便于书写的笔画系统了。"

"这就是说汉字的造字法和造字历史，说明汉字并不是先有笔画，再有部件，再造汉字的。汉字是先有了一些象形的'文'，这些'文'，又作为部件繁衍了众多的合体字。笔画系统形成在后，……正是从这个意义上说，我们认为部件拆字法符合汉字造字法的基本原理。……在这个