关于制定《汉语信息处理词汇》国家标准的若干问题刍议

张　普

制定《汉语信息处理词汇》的国家标准，是一个十分必要的课题，同时也是一个十分复杂和困难的课题，现在，这一课题已经提上议事日程，但是，要想顺利完成这个带有一定难度的课题，我们必须认真探讨制定这项标准的一系列有关问题，否则，这个"十分必要"的标准将会十分难产，这绝不是危言耸听。这就是本文论述的出发点。

一、关于制定标准的基本方针

我认为制定这一标准应采取积极慎重、多方协调、统筹规划、分集公布的方针。之所以要采取这种方针，是由《汉语信息处理词汇》标准的特殊性而决定的。汉语信息处理，或者叫中文信息处理(其本身就需要"正名")，是一门多边缘交叉的高技术，是与国家的四个现代化建设息息相关的带头学科。它既有自身的独立性，又与自然科学、社会科学的许多原有学科或新兴学科发生交叉，因此使该标准的制定既至关重要，也具有相当程度的复杂性，从前者(重要性)角度出发应当积极，从后者(复杂性)角度出发应当慎重。慎重表现为统筹规划、多方协调、分集公布等措施。

统筹规划和多方协调要从如下几方面进行：

1. **行政方面** 从语言文字规范化的角度看，这一标准的制定与国家语言文字工作委员会有涉；从工程应用的标准化的角度看，这一标准的制定与国家标准总局有涉；从统一自然科学术语的角度看，这一标准的制定还与全国自然科学名词审定委员会有涉。因之行政上需要三家协调，统筹规划，即使不能做到同一标准三方法定，也应力求三方的标准互不矛盾。

2. **学科方面** 陈力为同志在《中文信息处理的现状与展望》一文中曾经指出："中文信息处理的研究领域可粗分为二类：一是以计算机为工具对中文信息进行研究，二是在中文信息研究的基础上研究如何使计算机适应中文信息处理的要求。它是一个多学科密切结合的研究工作。"在学科方面，中文信息处理与信息科学、计算机科学、语言文字学、电子学、声学、心理学等都有密切联系，还牵涉到通讯技术、自动控制技术等领域，所以《汉语信息处理词汇》的许多内容将会与其他学科与工程技术方面的术语重合或交叉。在制定本标准时，选定规范术语，确定标准英译名，界定术语定义，都不得不同时考虑到相关学科的定名与定义，特别是务必与已正式公布的国家标准及ISO的世界标准协调。这种协调应是有条件的。根据什么条件协调和如何协调，这也是极棘手的研究课题。

3. **学术方面** 就中文信息处理学界内部而言，由于该学科是新兴的学科，并处于迅速

发展阶段，故大量名词术语尚未定型或统一，无论从科学性、稳定性、通用性中的任何一点来衡量，总有一些词语难于规范或限制。同时，从事中文信息处理研究的专家分别来自多种学科领域，对于同一客观事物会有基于各自学科领域的不同认识、不同分析，也就常常会给以不同的定称。即使命名一致，定义也不尽吻合。定名的分歧与定义的分歧往往引起学术上的争论，众说纷纭，莫衷一是。这是跨学科综合研究过程中的必然现象。这就给本标准在制定时的学术方面的协调带来了很大的麻烦。

4. **地域方面** 按照语言学的观点，方言才有地域性，同一民族语言中的科学术语应是超地域的。但是目前由于众所周知的原因，台湾省和香港、澳门地区使用的科学术语（不仅仅是中文信息处理方面的术语）与大陆不尽一致。在制定《汉语信息处理词汇》的国家标准时，我们应加强与台港地区同行的交流，取长补短，促进定名定义的统一。科学术语的统一于国于民都是有利的。最好是制定统一的标准，不能马上制定统一的标准时，也应迈出第一步，先统一名词术语，这是制定标准的基础。我们在审定标准时，应邀请台港地区的专家学者参加，同时还应征集使用汉字或汉语的其他国家（如日本、新加坡、南朝鲜）的专家学者的意见。在国际上，有人提出"汉字文化圈"的理论，这种理论是否成立或是否经得起考验另当别论，但加强使用汉语汉字的国家和地区在汉语汉字术语方面的协调和统一无疑对于信息处理技术是有益无害的。

二、关于制定标准的必要与可能

1. **必要性** 国家已将这一课题的研制提上了议事日程，并且建立了"《汉语信息处理词汇》国家标准研制工作组"。这正是因为看到了这一标准制定的必要性。我们从三方面概略论述必要性如下——

① 目前，中文信息处理方面的名词术语混乱、纷杂的现象较严重，给国际国内的学术交流带来了诸多不便。其大者如"中文信息"，多年来，中国中文信息学会对"中文信息"没有一致的解释，分歧的理解表现在一系列问题上：A."中文"是仅指书面语言（文字），还是包括有声语言在内？B."中文"是仅指汉语汉字还是包括中国境内少数民族语文在内？C."中文信息"是什么？"中文信息处理"是处理中国语言文字本身还是处理以中国语言文字作为载体的信息？D."中文信息处理"的英译名为"CHINESE INFORMATION PROCESSING"（见1983年北京中文信息国际研讨会），"中文信息"的英译名有时也作"CHINESE INFORMATION PROCESSING"而不作"CHINESE INFORMATION"（见《中国中文信息研究会章程》规定的学会英译名及学会杂志《中文信息》英译名）。如何解释英译时出现的不一致？同是中文信息学会的会员，但不同专业的人对"中文信息"却有上述种种方面的争论。小者如组成汉字的那些"构字部件"，它们的叫法五花八门，或谓"字根"，或谓"字元"、或谓"字素"，或谓"字母"、或谓"形母"、或谓"汉字图素"，传统也称为"偏旁"、"部首"等，均系异名而同实，其内涵与外延基本无别，只是命名的角度不同罢了，如据拼音文字词根组词的性质而将汉字构件单位称为"字根"，据自然科学中元素化合的性质将汉字构件单位称为"字素"或"字元"等等。总之，现行名词术语不统一的状况使得标准化的工作成为亟为迫切的任务（不统一的状况集中表现为名称不统一，定义不统一，英译名不统一三个方面）。

② 中文信息处理已从研究阶段进入推广应用阶段，因此与工程应用相关的一系列标准化问题日益突出，其中某些标准已经制定并颁布推行，例如：《信息交换用汉字编码字符

集》、《信息处理用点阵汉字字模集及数据集》(15×16, 24×24, 32×32点阵等)，另有一批标准正在筹划、论证、研制之中。《汉语信息处理词汇》作为国家标准应是有关中文信息处理的各项标准中最基本的标准，其中的《基本词汇集》又是基础之基础，首先制定《汉语信息处理词汇·基本词汇集》尤为必要。

③ 钱三强先生曾指出："目前，一门综合了信息科学与语言学，专门研究术语订名、概念、应用及其相互关系的新型学科——术语学的研究正在蓬勃开展，这门科学的研究水平已经成为发达国家科技水平的重要标志，根据术语学理论与研究成果，应用现代计算机技术建立起来的规模宏大的术语数据库已经在许多发达国家建立起来，在社会生活的各个方面发挥着重大作用。而术语数据库的建立也依赖于名词术语的统一与定义的确切。由此可见，自然科学名词术语的统一不仅是国家在科学文化方面的一项基本建设，也是新技术革命对我们的紧迫要求。"① 术语学是"综合了信息科学与语言学"的一门新型学科，术语数据库的建立又要"应用现代计算机技术"，而中文信息处理恰恰主要是信息科学、语言学和计算机技术的密切结合的高技术研究，它理应也必须在名词术语的统一化和标准化方面先行一步，为建立我国的术语数据库奠定基础。

④ 随着中文信息处理在应用方面的推广，从事该项研究工作的队伍不断扩大。现有的队伍大多是各种学科（如计算机科学专业、信息学专业、语言学专业等）的人才横移所形成的，而目前综合人才的培养已提上议事日程，有关中文信息处理的研究室、研究所、研究中心、系科甚至专门化学院相继问世或酝酿组建，相应的专著和教材也陆续编写出版(包括公开发行或非正式发行者)。中文信息处理名词术语不统一和无标准给教材编写与教学工作均造成许多不便。

以上是对制定《汉语信息处理词汇》国家标准的必要性的论述。

2. **可能性** 首先，中文信息处理作为独立的新兴学科已日趋成熟，中文信息处理工作者对该学科的研究内容、研究方法、学科体系及它与其他学科的关系逐步有了更深入的认识。近年来，中文信息处理在基础理论和应用技术两个方面都有长足的进展，因而使我们有可能对曾经有争议的某些问题重新考析，实现认识上的统一。中文信息处理名词术语的混杂是历史发展过程中自然形成的(虽然这"历史"还非常短暂)，而今天则到了应当统一也可能统一的阶段。"中国中文信息研究会"更名为"中国中文信息学会"，即基于中文信息处理学科已成长，壮大的事实，因此可以认为名词术语的标准化具备了一定的学术基础。

其次，我们目前已拥有足够的资料(相对而言)来进行该项标准化研究。截至1986年中国中文信息学会全国会员代表大会止，仅中国中文信息学会已召开全国性学术会议25次，入选论文1300余篇。如果加上各地方分会的学术交流会，加上其他兄弟学会(如中国仪器仪表学会汉字信息处理系统研究会、中国计算机学会人工智能研究会等)所发表的有关中文信息处理方面的论文，特别是加上近几年来在国内或国外召开的有关国际会议的论文，资料更为丰富。即是说，标准化工作具备了一定的资料基础。

第三，与《汉语信息处理词汇》相关的某些国家标准，有的已研制多年，有的已正式颁布，如《数据处理词汇》的国家标准已研制了十多个子集，并且翻译了国际标准作为参照。因此，在制定《汉语信息处理词汇》国家标准时，特别是处理那些与其他标准相关或交叉的条目时就有了较好的参照基础。

第四，标准采取分集制定的方法。首先制定基本词汇集。这样可使某些确实还处于初期发展阶段的子集或一时尚难于标准化的子集在稍后的时间内制定。

三、关于制定标准的难点

1. 与相关标准的关系

中文信息处理既然是交叉学科，它的词汇不可避免地会与其他有关学科的词汇交叉。这些交叉条目如何定义？当其定义与有关学科交叉条目的定义有杆格之处时或不尽相同时，应如何处理？这些问题均应确立适当原则。例如，在《数据处理词汇》中对"语言"、"自然语言"与"汉字"三条目的界说如下：

语言：language为了传递信息而使用的一组字符、约定和规则。

自然语言：natural language一种语言，其规则是根据当前流行的用法而不是明确的形式规定。

汉字：ideogram；ideographic character自然语言中的一种图形字符，它表示一个事物，概念及与之相联系的发音。

例：中国汉字或日本汉字。

这里有两类关系需要协调：一是条目相重时怎样处置，二是定义与英译名不相符或有矛盾时怎样处置。要协调两类关系必须首先明确《汉语信息处理词汇》的性质及适用范围。前举《数据处理词汇》规定的适用范围是："有关电子计算机及信息处理各个领域的设计、生产、使用、维护、管理、科研、教学和出版等方面。"(着重号系引者所加)我们是将《汉语信息处理词汇》视为《数据处理词汇》中信息处理范畴的一个子集，抑或视为平行的独立的国家标准？只有这一前提解决了，我们方能进一步确定《汉语信息处理词汇》本身的分集原则、收条原则、界说原则等一整套细则，这样才不至于与既有的相关标准条目发生大面积重复，同时又保持了自己的独立性和系统性。

2. 与ISO国际标准的关系

汉语和汉字信息处理的标准理所当然应该由中国、由中国人来制定，这是我们责无旁贷的任务。而国家标准一经确立，即应报国际标准化组织争取享有国际标准的地位。因此，国家标准的制定必须有长远考虑，考虑与ISO国际标准的关系。凡ISO其他标准中已定义的条目尽量采用，采用的方式(是直接采用还是等效采用还是其他方式?)也应认真研究，要以能为ISO接受为国际标准为原则。

3. 英译名的处理

许多科学名词术语常是外来词，它们本来已有标准的或通用的英文名称，甚至已有国际标准可循，汉语中的称呼只不过是其中文译名。而《汉语信息处理词汇》中大量名词术语是汉语特有的，无对应的现成英文名称，如何将它们准确地译成英文(不仿害汉语中的原意)则相当复杂。前述"中文信息"与"中文信息处理"两词语的英文译名失去区别、混为一体就是一例，其他如：笔画、笔顺、偏旁、部首等的英译都十分困难。习惯沿用的译法中有的不科学，有的有出入。如何协调作为标准的定称与传统译法的关系也绝非易事。这方面尤其需要海外华人和汉学家(特别是中文信息处理专家)的建议和评价。

四、关于标准的分集和其他

1. 关于分集：

我们建议《汉语信息处理词汇》分如下11个子集：

① 基本词汇
② 汉语和汉字
③ 汉字编码
④ 汉字字符识别
⑤ 汉语言语识别与合成
⑥ 汉语理解
⑦ 汉语的机器翻译
⑧ 汉语信息处理设备
⑨ 汉语信息处理软件
⑩ 汉语信息处理应用技术
⑪ 其他

首先要制定的是基本词汇集，基本词汇集的收条原则要考虑到它与其余各集的关系。

2. 关于建立中文信息处理资料库

制定《汉语信息处理词汇》国家标准的工作可以与建立"中文信息处理资料库"的工作同步进行。资料库的内容应包括文献资料(含论文、专著、专集、期刊等)及档案资料(包括各研究机构、学会组织机构、国际国内学术会议等)。这种资料库，一方面可以作为建立中文信息处理的情报检索系统的基础，实现全文检索、提要检索、标题检索等，另一方面，只要稍加改造或者录入时增加一些标志符，就可以为制定《汉语信息处理词汇》提供自动收条、自动筛选条目等方便，也可以为推敲定义与英译名提供大量的资料基础。在此基础上，我们还可以实现《汉语信息处理词汇》国家标准的计算机辅助审订和修订工作，从而大大提高国家标准的审订与修订的效率。如果实现了国家标准的编、审、修的计算机化，这套系统软件无疑具有广泛的推广价值。

3. 关于建立中文信息处理术语数据库

建立"中文信息处理资料库"是编制《汉语信息处理词汇》的必要前提，而建立"中文信息处理术语数据库"是编制《汉语信息处理词汇》的必然结果。我们在前文引用的钱三强先生《统一自然科学名词术语的意义重大》一文已经指出："根据术语学理论与研究成果，应用现代计算机技术建立起来的规模宏大的术语数据库已经在许多发达国家建立起来，在社会生活的各个方面发挥着重大作用。而术语数据库的建立也依赖于名词术语的统一与定义的确切。"显然，一旦《汉语信息处理词汇》的国家标准制定颁行，术语将会统一，定义将会明确，译名将会规范，那么"术语数据库"的建立也就水到渠成了。把《汉语信息处理词汇》的国家标准研制工作与建立相应的"术语数据库"的工作结合起来进行，无论怎样看都是十分有利的。

《汉语信息处理词汇》的制定是中文信息处理界的一件大事，希望这件我们自己学术领域中的大事得到海内外学者的关注和支持。

注释：

① 钱三强：《统一自然科学名词术语的意义重大》，见《人民日报》1985年6月20日。