(云南师范大学学报（哲社版）》2008 年第 6 期)

标准、平台统一与资源整合——多语言知识库建设的思考和建议

王铁琨

“多语言知识库”是教育部语言文字信息管理司立项支持的民族语言文字标准化、信息化重大项目，也是“中国语言资源数据库”建设和“民族语言文字信息化平台”的重要组成部分。研讨和组织建设多语言知识库，是一个功在当代、利及子孙、造福国家民族的系统工程。但是这样一个库的建库原则是什么，库里应该包含什么样的基本内容，建库的时机、步骤、运作机制和分工，以及经费筹措等等问题，需要学者专家多方面、多角度的研究和语言文字工作部门审慎、科学的决策。从这个角度看，召开第一届多语言知识库研讨会非常必要。

当今全世界都在关注各种语言资源的保护和开发、利用，与之有关联的是联合国设立“国际语言年”和“国际母语日”。联合国确定 2008 年为“国际语言年”，并连续九次庆祝“国际母语日”，主要是为解决随着信息技术的快速发展所带来的各民族语言发展的不平衡问题和日益加剧的“数字鸿沟”，把各民族的语言、文化和发展有机地融合在一起，进而促进语言的多样性和文化的多样性，共同建设和谐世界。我们认同“国际语言年”的主旨精神，当然支持符合这样主旨精神的第一届多语言知识库研讨会。

今年 4 月 11 日，在北京举行的联合国“2008 年国际语言年暨第九届国际母语日”庆祝活动开幕式上，我曾代表中国语言文字工作部门讲过这样一番话：

中国是一个统一的多民族国家。中国要和平发展，必须推广使用一种各民族共同接受的语言文字，即普通话和规范汉字。与此同时，中国通过立法和一系列语言规划，科学处理语言文字主体性和多样性的关系，推动国家通用语言文字和少数民族语言文字的学习、使用和共同发展，中国的《宪法》《民族区域自治法》 和《国家通用语言文字法》都明确规定“各民族都有使用和发展自己语言文字的自由”。

由于上述语言政策和法律的贯彻实施，中国的语言文字工作取得了世人瞩目的成绩。但是在世界经济一体化和中国城镇化进程不断加快的新形势下，我们也注意到，由于强势语言的影响，语言多样性和文化多样性正面临着严峻的挑战。当前，解决语言问题、保护语言资源和维护语言权利等方面的工作尤其重要。

中华人民共和国成立以来，政府在解决语言沟通、语言压力和语言濒危等问题上做了大

量的工作，成效显著。但是，仍有许多新的矛盾和问题亟待解决， 如正确处理普通话与方言的关系、 各民族语言文字之间的关系、母语与外语的关系，等等。

语言不仅是信息的载体和交流的工具，语言更是重要的、不可再生的文化资源。当今人类语言资源正在快速流失，这与生物物种的流失和大气变暖等同样值得关注。基于珍爱中华语言资源的理念，中国正在酝酿开展新世纪的语言普查（后更名为“中国语言资源有声数据库建设”，作者注），以期建立可永久保存的中国语言多媒体语料库及相关数据库， 绘制详细、准确、可传至后代的多媒体语言地图，建立需要保护的语言、方言目录，开发和利用好国家语言资源。我们愿意和全世界的同行们携起手来，共同保护人类的语言资源。

语言权利包括个人和群体的语言权利，涉及公民的生存权和发展权。应该采取有效措施，切实维护和保障公民和群体的语言权利， 如母语学习权、母语使用权和母语研究权，以及获得各种语言服务的权利。

这几段话中，最重要有三点：解决语言问题、保护语言资源、维护语言权利。

这三方面工作，是当今全世界的语言工作者共同关心的问题。而我们现在研讨的“多语言知识库”，既涉及语言问题的解决，也涉及语言资源的保护，更涉及语言权利的维护。所以，提出“多语言知识库”这个命题很有价值。

近年来，教育部、国家语委提倡树立“珍爱中华语言资源，构建和谐语言生活”的全新理念，这一理念是在工作实践中逐步形成的。语言是信息和文化的载体，是思维和交际的工具，这样来认识语言当然没有错。但语言更是一种文化资源，是同森林资源、矿产资源、水资源一样重要的不可再生的国家资源。如果仅仅把语言看作“载体”和“工具”，那么全世界有一种语言文字就足够了。而把丰富多彩的语言文字看作人类共同的财富和资源，就会更加珍爱它、保护它、开发它、利用它；就不会只关注语言分歧、语言压力所带来的一些矛盾和社会问题，而是在解决这些矛盾和问题的过程中重在建设、重在服务，科学处理语言文字主体性和多样性的关系，充分发挥语言资源在国家和平发展和走向世界中的作用，营造健康和谐、多言多语的语言生活。

基于上述理念，2004 年以来，教育部语言文字信息管理司与有关高校和行政主管部门合作，陆续组建了国家语言资源监测与研究中心的平面媒体语言分中心（设在北京语言大学）、有声媒体语言分中心（设在中国传媒大学）、网络媒体语言分中心（设在华中师范大学）、教育教材语言分中心（设在厦门大学）、海外华语研究中心（设在暨南大学）和少数民族语言分中心（设在中央民族大学），并通过各分中心建设的动态流通语料库，有计划地开展语言使用实态的考察与研究，范围几乎涉及到中国各种语言文字及各个应用领域、层面。

监测研究的相关成果和数据，以“中国语言生活绿皮书”和《中国语言生活状况报告》的形式定期向社会发布。这一举措，使国家语言资源得到了一定程度的保护和开发，使社会语言生活得到了科学、健康的引导，监测、研究成果在国内外学术界和社会上均产生了积极的反响。

在民族语言文字资源库建设方面，除国家语言资源监测与研究中心少数民族语言分中心正在建设相应的动态流通语料库外，教育部语言文字信息管理司还资助有关高校陆续建设了藏文、蒙古文、维吾尔文和朝鲜文等语种的语料库，从而使得一些传统通用的少数民族语言在资源库建设方面有了初步的基础，开始形成语言资源建设的较为完整的系列。当然，同国家通用语言文字的资源库建设相比，民族语言文字资源库建设由于起步较晚，无论从技术基础和经费投入来看，差距仍然比较大，远远不能满足需求。

民族语言文字资源库（包括“多语言知识库”）建设还受到标准、平台和资源等方面一些因素的制约。这三个方面是建库的前提，必须着力解决好。

一是统一标准。“多语言知识库”建设首先要坚持统一标准，否则无法做到兼容和共享。这个标准就是国际标准，以及在 ISO/IEC 10646 框架下的国家标准。随着信息化的发展，现在整个地球都成了一个“村”，我们在标准建设上也不能搞“窄轨铁路”，自我封闭。在建库步骤上，可以考虑已有国际标准的语种先建，从“双语”开始，逐渐形成“多语”。坚持走国际标准化框架下的国家标准化道路，不但适用于中国独有的语言文字，更适用于跨境的民族语言文字。比如朝鲜语语料库，是按朝鲜民主主义人民共和国的标准建，还是按韩国的标准建，抑或是按照我国朝鲜族聚居的延边朝鲜族自治州的标准建，我想首先还是要看该语言文字是否已经形成了国际标准。如果已有现成的国际标准，最明智的办法当然是“等效采用”；若没有，则可“以我为主”，通过国际标准化组织争取我们的标准率先成为国际标准。总之，这个问题需要用世界眼光来认真思考和决策。

以上谈的是标准的大的方面。数据库、知识库建设中标准、规范无处不在，其中包括选材（资源选择）规范、资料整理规范、文本录入规范、校对规范和建库规范等一系列内部的规范和标准，都需要先行统一。否则，各行其是，各自为政，库建起来也发挥不了预期的效用，甚至可能花钱“打了水漂”。

二是统一平台。“多语言知识库”建设要坚持统一平台。信息技术产品和科研成果，由于平台不统一所带给人们的困惑和烦恼，相信大家都遇到过。同一语种不同省区采用不同的平台，甚至同一个研究院所不同研究人员建立的语料库也不能打通使用，电子出版系统的文本资源需要通过转换软件的“帮忙”才能在信息处理系统实现共享，凡此种种，不利于国家

的信息化建设。中国各民族文字多种多样，有纯表音的，有表意的，有表音兼表意的，有罗马字母式的，有汉字系的（如古壮文和西夏文），有的从左至右书写，有的从右至左书写，有的自上而下书写……现在信息技术发展越来越快，如何采用必要的技术手段，使上述不同体系、不同书写方式的文字在一个统一的平台上实现兼容和共享，应该是能够做到的。正在建设中的“中华大字符集”就要设法解决这个问题，“多语言知识库”在总体设计上也要对此先行思考和抉择。我们有那么多胸怀远大、水平不错专家，相信“多语言知识库”的平台建设问题也能得到圆满解决。否则，缺乏全局视野和国际眼光，“多语言知识库”乃至“国家多语种语言资源平台”难以真正发挥“信息共享”和“资源保护”的功能。

三是资源整合。语言资源有语种之分（如汉语、藏语、蒙语、维吾尔语等），有地域之分（民族语言也有各种方言，如藏语就有三大方言），有古今之分（如蒙文有现代蒙古文与古八思巴文），有境内境外之分（许多跨境语言文字如傣文、苗文、哈萨克文，境内外有程度不同的区别），从载体上又可以区分为平面媒体语言、有声媒体语言和网络媒体语言。如何科学地采集和整合这些重要的语言资源，也需要研究和规划。

语言资源不存在统一不统一的问题，而有一个是否科学、合理和适用的问题，自然，采集语言资源的内部规范也需要统一。目的不同，资源库建设的类型和内容也会不同，比如，有服务于语言资源监测研究的(如国家语言资源监测与研究中心各分中心建设的动态流通语料库），有服务于词典编纂的（如双语词典、多语词典和各种学习型辞书编纂所建立的大型语料库），有用于进行语言资源历时比较的，有用于进行语言资源共时研究的，等等。资源库建设还有一个整合已有资源和开发新的资源的问题，力避重复建设。比如，国家语委目前正在准备启动“中国语言资源有声数据库建设”，规划中全国各县市（包括各民族地区的县市）都会有一些采集点，“多语言知识库”作为“国家多语种语言资源平台”的重要组成部分，可否与之相结合，使花费很大气力、大量资金采集的语言资源能够尽可能多地发挥出社会效益？这些都属于资源整合问题，需要统筹规划，认真研究决策。总之，路要一步一步脚踏实地地往前走，不能急于求成！

以上结合“多语言知识库”这个命题谈了个人一些思考和建议，其背后涉及宏观统筹、学术梯队建设以及经费投入等许多深层次的问题，希望能够引起大家的讨论和批评。