DOI：10.3969/j.issn.1672-8874.2011.02.039

信息化条件下的语言资源建设与应用

邢富坤
（解放军外国语学院，河南 洛阳 471003）

［摘 要］ 信息技术在语言教学领域的广泛使用对语言资源建设提供了新的机遇，同时也提出了新的挑战。在分析已有语言资源建设存在的问题基础上，提出了语言资源建设要遵循的主要原则，并描述了以互联网新闻资源为基础的语言资源建设工作。

［关键词］ 信息化建设；语言教学；语言技术；语言资源
［中图分类号］ G434 ［文献标识码］ A ［文章编号］ 1672-8874（2011）02-0118-03

语言教学需要真实、地道、鲜活的语言资源支持，随着信息技术的迅猛发展及其在语言教学领域的广泛应用，语言资源建设面临着新的机遇与挑战。互联网的出现为语言资源建设提供了宝贵的机遇，国外一些研究者已经试图挖掘互联网的潜在语言资源价值，面向语言教学建设语言资源库。但信息技术的发展也对语言资源建设提出了挑战，如何有效利用存在于互联网上的语言资源，提高语言资源的建设效益成为当前亟待解决的问题。

本文首先对语言资源进行分类讨论，然后分析已有语言资源建设中存在的问题，并提出语言资源建设要遵循的主要原则，最后介绍了我们自主实现的以互联网新闻资源为基础的语言资源建设系统，并对未来工作进行了初步展望。

一、语言资源的分类

由于目前绝大部分语言资源建设都依赖于互联网开展，因此我们根据存在于互联网上的语言资源的生产时间和生产者特征将其分为三类：

（一）互联网出现前的语言资源

这类资源的年代一般较为久远，并且在其数字化之前，一般都是以纸质媒体为存储和传播媒介。当这类语言资源经过数字化处理，并以网络为媒介进行存储、传播和使用后，就成为基于网络的语言资源。这类资源的主要特征是语言较为规范，内容较为可靠，经过数字化处理后便于计算机存储和使用，但是其规模有限，语言缺乏时代性。

（二）互联网出现后且由专门媒体机构生产的语言资源

这类资源包括如报纸、期刊、广播、电视等传统媒体机构，通过建立网站而使得媒体内容在网络上存储、传播和使用。这类资源的主要特点是语言鲜活真实，内容可信度较高，具有较快的更新速度，同时由于这类资源具有了数字化形式，比传统媒体形式的资源更便于计算机存储和处理。

（三）互联网出现后且由普通网络用户生产的语言资源

这类资源是互联网出现后发展极快的一类语言资源，并日益成为网络资源的主体，其代表形式主要包括邮件列表、新闻组、论坛、博客、微博、维基百科等。这类资源的主要特征是用户直接参与网络信息的创造，其参与范围不再局限于少数人或机构，由于参与的广泛性使其具有较大的规模，并覆盖多种语言，具有传统媒体不具备的及时更新性。

综合考虑以上三类语言资源的特征，第二类语言资源在获取难度和使用效果上更具有优势，因此目前工作主要集中于此类语言资源。而第一类资源虽然获取难度不大，但由于其规模一定，内容具有封闭性，因此本文对其不做过多讨论；第三类资源由于其语言内容的可靠性和规范性难以得到保证，因此目前并不将其作为语言资源建设的主要来源。但第一与第三类资源都对语言教学都有着重要价值，第一类语言资源是传统的语言教学素材，对于这类资源应着重于其内容的深度加工和使用；而第三类语言资源由于其语言的鲜活性和时代性也对于语言学习有着重要帮助作用，对于这类资源应着重于资源的获取和筛选。

二、语言资源建设中的主要问题

尽管语言资源建设一直是语言教学中重点关注的对象，但是由于受限于技术手段，语言资源建设中还存在较多问题，主要包括以下两个方面。

（一）语言资源的获取、加工主要依靠人力，严重制约了资源的建设规模和使用效益

尽管互联网的发展为语言资源建设提供了丰富的基础材料，但由于互联网上的信息类型多样、信息内容繁杂且分布较为分散，将这些基础材料转化为语言教学资源还需要经过获取、评估、整理、加工、保存等一系列环节。目前这些环节的工作主要是依靠语言教学者来完成，而教学者在资源建设中的大部分时间花费在了浏览网页，复制粘贴等简单重复性劳动上，而在语言素材评估、加工和编辑等上面的时间受到很大影响。此外，由于教学者的时间精

［收稿日期］ 2011-06-20
［作者简介］ 邢富坤（1975-），男，河北唐山人，解放军外国语学院基础部信息技术教研室讲师，博士。

力有限，在有限的时间内不可能浏览太多网站，更不可能及时跟踪各个网站上的更新资源，这就使得语言资源建设规模受到较大限制，难以为语言教学提供全面且及时更新的语言素材。

（二）语言资源建设的语种分布不平衡，加工程度较浅

目前语言资源建设的重点主要集中于英语等少数通用语种，而非通用语种的语言资源则较为匮乏。我们考察了国内几家主要的语言学习网站，发现绝大部分只提供英语学习材料，而少数网站，提供法语、西班牙语、日语、韩语等语种的学习材料，但是材料的规模远小于英语学习材料，而对于像印地语、尼泊尔语、斯瓦西里语等非通用语种的语言资源基本没有。同时，我们也发现在这些网站上的英语学习资源绝大部分都还是原始文本，基本没有经过任何语言层面的加工和处理，例如没有提供词频、句频、词汇等级、重点词释义、专有名词标注及解释等信息，这些不足都使得语言资源的自身价值和服务教学的效益受到制约。

三、语言资源建设与使用的主要原则

语言资源的建设与使用具有其内在的规律，总结语言资源建设与使用的主要原则有以下几点。

（一）静态与动态结合，以动态为主

语言资源可以分为静态资源与动态资源两种，静态资源是指已经固定不变的语言素材，经典的文学作品是典型的静态资源，而动态语言资源则是处于更新变化中的语言素材，实时更新的新闻报道是典型的动态资源。静态资源与动态资源各有其特点与优势，因此在语言资源建设中，要兼顾静态与动态，不能偏废。由于新鲜及时的语言素材能够较好地激发语言学习者的学习积极性，同时能够为语言学习带来最新的语言知识，使得学习者习得的语言能够紧跟语言的发展趋势，因此要将动态语言资源建设作为语言资源建设的主要内容，实现对语言资源的定期更新，从而使得学习者能够及时接触到目标语的语言材料，在语言使用中提高对目标语言的掌握能力。

（二）人工与自动结合，各取所长

语言资源的建设途径有人工和自动两种，传统的语言资源建设一般都是由人工进行，由于人力限制，使得语言资源建设的速度和效率都难以得到保证。自动构建语言资源虽然能够大大加快建设速度，提高建设效率，但是由于机器并不知道什么样的语言素材适合作为语言资源使用，因此自动构建也存在困难。解决这种困难的途径就是将人工和机器有机地结合起来，发挥各自不同的优势，实现人机互助的语言资源建设模式。

表1 人与机器的特点对比

	任务分析	抽象概括	算法设计	存储	匹配	替换	计数
人	+	+	+	?	?	?	?
机器	-	-	-	+	+	+	+

注：“+”表示能完成且擅长，“-”表示不能完成，“?”表示能完成但不擅长。

从上表可以看出，人与机器在不同的任务中各有不同的表现，因此在语言资源建设中，必须正确处理好人与机器的关系，让二者扬长避短，最大程度地发挥各自优势，促进语言资源的建设工作。对于人而言，其任务是需要对语言资源的概念、资源的来源、获取的方式、处理的方法等进行明确，形成形式化的表达方式，进而形成有效的机器指令。同时，人还应对机器的任务完成情况进行监督和评价，及时发现问题，提出解决办法。对于机器而言，则尽可能地发挥其存储、查找、匹配等能力优势，在资源访问、目标抽取、标记过滤、语料入库、信息检索等任务中发挥作用。

（三）建设与使用结合，需求引领

语言资源的建设是为了使用，建设与使用必须有机结合在一起，防止只建设不使用，或者是脱离使用，盲目建设。语言资源的使用决定了要面向不同的语言使用者开展不同层面的加工工作，从而为不同的对象提供更加适合的语言材料。对于学习者而言，要对语言素材进行难度上的区分，为不同语言水平的学习者提供不同难度等级的语言材料，同时要对语言材料中的特定词、表达方式以及专有名词等做出标注和解释，从而方便学习者的使用；对于教学者而言，要能够给出语言材料的来源、作者等描述信息，从而方便教学者确定语言材料的可靠性与权威性，同时还要给出语言材料的词汇范围，最好与现有语言教材进行对比分析，以便教学者确定语言材料的适用性。

四、语言资源建设的主要环节

语言资源的建设主要包括语言资源的获取、加工、管理和使用四个主要环节，在这四个环节中人和机器发挥各自不同的作用，承担不同的任务。

（一）语言资源的获取

语言资源的获取是指机器根据人指定的目标站点和预设的获取条件对特定资源开展获取工作。语言资源获取的重要前提是目标站点的确定，对于新闻语言资源获取而言，目标站点是主要新闻媒体机构的官方网站。选择目标站点是人必须承担的任务，而不能由机器自动完成，人可以利用丰富的经验和综合知识对各类网站进行评价和选择，将那些内容丰富、信息可靠、更新及时、语言规范的网站纳入到自动获取的范围之列，而对于不符合这些要求的网站则暂不予考虑。

（二）语言资源的加工

根据加工深度的不同，可分为网页格式标记处理与语言深度加工两个层面。网页格式标记处理可以通过机器自动进行，人只需提供网页格式等模板信息即可。而语言层面的加工则需要人机的深入合作，同时语言加工也是语言资源加工的重点所在。语言层面的加工目标是为获取的语言资源标注更多的语言学信息，以便能够为教学使用提供更大便利。语言层面的加工工作主要包括词、语句、篇章三个层次，词层面的加工包括词频统计、不同难度等级词的分布统计、关键词的自动识别和释义，人名、地名、机构名等的自动识别和标注；语句层面的加工包括自动分句、句数统计、句长统计、句法自动分析、语义角色自动标注等；篇章层次的加工包括语篇难度等级的测量、语篇的自动翻译、语篇的自动分类、自动摘要或简写等。随着加工

层次的递升，加工的难度也逐级增加，需要人介入的程度也越来越大。

（三）语言资源的组织管理

语言资源的组织管理直接关系到使用效率，良好的组织管理方式能够为特定资源的查找检索提供便利，使得用户能够在较短时间内，较为准确地查找到目标内容。数据管理模式要能够满足用户通过日期、语种、媒体、关键词等为查询字段对语言资源库进行查询，并保证查询的效率和准确性。对于大规模文本类型语言资源的查询，一般采用建立索引文件的方式进行。建立索引文件后，能够在较短时间内完成用户的检索要求，但是建立索引的过程较为长，且索引文件占用的存储空间较大，更新较为复杂。本研究将获取和加工后的语言资源存储为数据表的格式，数据表中包含了语言资源的主要信息，如报道时间、媒体、语种、外文标题、外文内容、中文标题、中文内容、新闻长度、新闻类别等信息，然后利用数据库查询语言（SQL语言）对数据库进行查询。这种数据管理方式的最大优点是利用了数据管理系统的优势，可以较为方便的进行管理、修改、更新和查询，尤其重要的是可以较为方便地利用网络平台对采集后的数据进行发布，方便用户使用。

（四）语言资源的使用

语言资源建设的主要目的就是为了能够更好地被语言教学者和学习者使用，使得语言资源的效益得到发挥，因此使用是语言资源建设的主要目标。语言资源的使用可以分为本地下载使用和网络访问使用两种。本地下载使用是指用户将语言资源下载到本地计算机上，然后用户根据需要使用。这种资源使用方法面临两个主要问题，一是下载语言资源需要占用用户的时间和存储空间，并且用户还需下载使用语言资源的工具，并保证这些工具能够在本机上正常运行，这些都增加了用户的使用负担，妨碍了资源的有效使用；二是由于语言资源建设需要较多投入，很多资源建设者并不希望语言资源被用户完全获取，而是希望用户根据需求有选择地使用资源。另外一种使用方式是语言资源通过网络进行发布，用户以网络为媒介，通过访问语言资源所在站点来使用资源。这种方式既保护了语言资源的所有权，也满足了用户对语言资源使用的需求，并且语言资源的开发者可以根据用户需求开发相应的应用功能，用户可以通过浏览器直接使用这些功能，而不需要额外安装其他工具，这使得语言资源的使用更加有效。在本系统中使用了网络访问的资源使用方式。

五、语言资源建设工作介绍

针对语言教学的需求和语言资源建设的主要原则和方法，我们自主构建了一个基于互联网资源的多语种、多来源的动态语言资源系统，这里对获取的资源和使用情况简要介绍。

（一）素材来源

目前本系统的目标媒体来源主要包括BBC、CNN、路透社网、新华网、中国日报网、中国国际广播电台网等6家新闻媒体机构的官方网站，这些网站可以被分为美洲、欧洲、中国三大区域，而各网站内部又可以按照不同语种分为不同类别。选择不同区域的有代表性的权威新闻媒体网站作为目标网站，收集不同区域网站上的新闻内容，有利于在语言教学中对比分析同一种语言在不同地理区域的使用特征，也有利于语言的对比分析和研究。例如，收集CNN、BBC和China Daily的新闻资源，可以很好地为对比研究美国英语、英国英语和中国英语的特点提供支持，从而有助于不同区域的英语教学和学习。

（二）语种分布

目前本系统能够对包括汉语在内的23个语种的新闻语料进行获取，具体语种有英语、日语、朝语、俄语、德语、法语、西班牙、葡萄牙语、阿拉伯语、土耳其语、乌克兰语、越南语、老挝语、泰语、印地语、普什图语、希腊语、印尼语、乌尔都语、斯瓦西里语、孟加拉语等。在这些语种中既有通用语种，也有非通用语种，其地理区域覆盖了世界主要大洲。

（三）加工内容

加工主要集中在英语新闻语料上，分别在词、句和篇章三个层面对英语语料进行了加工，而其他语种则主要是在词面做了粗布加工工作。主要加工工作包括自动分词及词频统计、学术词汇的自动标注和释义、人名、地名的自动识别、自动分句和句频统计、语篇的自动翻译等。

（四）未来工作

未来工作重点将放在对获取语料的深度加工上，尤其是语言层面的加工工作。主要工作包括更多语种的语言加工、文本的自动分类、文本难度测量、试题开发、多语种互动等工作。

初步建成的语言资源库已经在本院内部开始测试使用，使用效果较好，尤其是受到非通用语种的教学者和学习者的普遍欢迎，已经成为这些语种教学和学习的重要资源。

[参考文献]

[1] Emiliano Guevara. Proceedings of the Sixth Web as Corpus Workshop[C]. NoWaC: a large web - based corpus for Norwegian. Los Angels, 2010:1 - 7.

[2] George L. Dillon. Proceedings of the Sixth Web as Corpus Workshop[C]. Building Webcorpora of Academic Prose with BootCaT. Los Angels, 2010:24 - 16.

[3] 王建新. 计算机语料库的建设与应用[M]. 北京:清华大学出版社,2005:203.

[4] 黄晓斌. 网络信息挖掘[M]. 北京:电子工业出版社,2005:99.

[5] Christopher Manning & Hinrich Schutze. Foundations of Statistical Natural Language Processing[M]. Massachusetts: MIT Press, 2005:75.

（责任编辑：洪巧红）