《汉语语言与计算机学报》2003 年第 2 期，新加坡

编者的话

特约编辑　冯志伟

本期《汉语语言与计算机学报》是特刊。主编赖金锭教授和俞士汶教授邀请我当特约编辑，并且要我以特约编辑的名义写几句话。于是，我从命作文，写了这篇短文，算做是“编者的话”（write-up）。

2001 年 11 月 27-29 日在新加坡召开了 ICCL2001 国际会议 ，本期的文章都是从这次会议的论文中选出来的。ICCC2001 国际会议的主题是“Chinese e-Learning in the New Millennium”，“Chinese e-Learning”所包括的领域是非常广泛的，其中，网络上中文资源的自动获取、计算机辅助汉语教学以及基于 Internet 的远程汉语教学，都是”Chinese e-Learning”的重要内容，这些研究都与语料库的建设以及语言资源的开发有着密切的关系。本期的文章就是围绕着这些内容来选取的。

《网页信息的抽取与集成》以人们十分关心的招聘网站为例子，介绍了中文网页信息抽取与集成的方法以及作者研制的系统。整个系统包括预处理、信息抽取和信息规范化等三个部分。该系统抽取网页信息的平均准确率、召回率和 F 值分别为 75.51%、56.48%和 64.62%，性能比较好。应该说，作者在 Chinese e-Learning 方面取得了很好的成绩。

《大规模标注汉语语料库开发的基本经验》专门讨论语料库建设问题，这是 Chinese e-Learning 的基础性工作。北京大学计算语言学研究所与富士通公司合作正在开发规模达 2,600 万汉字的《人民日报》标注语料库。标注语料库的工作要把文本中的句子按词语切分开，并且标注上词性等标记（标记集中约有 40 个不同的标记）。截止 2001 年 6 月底，已完成 1,800 万字的工作量。本文主要介绍他们在开发如此大规模的标注语料库过程中所积累的经验，涉及到规范的制订、辅助软件的开发以及工作流程的安排等问题。同时也介绍日本学者田中康仁整理的《红旗》杂志语料库。

《一个用于对外汉语教学的课件资源积累与提取模型》讨论的是计算机辅助汉语教学以及基于 Internet 的远程汉语教学中的课件建设问题。这显然是 Chinese e-Learning 非常重要的内容。在对外汉语教学中，需要大量的可以实现自组织积累和开放式共享的汉语课件资源。这些课件所需要的教学资源的获取形式以及积累形式，对于课件资源的合理利用有着非常重要的影响。本文从这两方面入手，阐述了共享数据库与课件数据库共存的必要性以及它们之间的通讯，描述了课件生成与这两种数据库之间的关系。最后进一步给出了本文中提出的课件资源模型方法在对外汉语教学课件中的一个应用实例。

《确定切词单位的某些语法因素》专门讨论了语料库自动切分中切分单位的确定问题。这是在中文信息处理中长期议而不决的困难问题，也是 Chinese e-Learning 研究中的一个基本理论问题。本文指出，传统语言学对于语素与词的划分存在着逻辑上的矛盾，这种矛盾导致在语素和词之间产生一个交集，这个交集，从语素的角度看是自由语素，从词的角度看是单纯词。由于自由语素和单纯词名异而实同，致使合成词和词组之间的界限不清，造成了划水难分的困难局面。这是汉语文本切分单位的确定长期议而不决的症结所在，也是传统语言学理论上的一个严重缺陷。为了克服这样的缺陷，本文提出了信息词(NLP word) 的概念，系统地研究了确定信息词的语法因素，分析了基于语法因素的替代测定法、插入测定法、黏附性测定法、功能完备性测定法等在确定切词单位中的效用。本文在 ICCC2001 会议上发表时，引起了与会者的热烈讨论，获得最佳论文奖。

《汉字义类信息库的研究与实现》讨论汉字义类信息库的建设问题。汉字义类信息是 Chinese e-Learning 的重要语言文字资源，汉字义类信息库的目的在于建立覆盖国家标准 GB-2312 所包括的全部汉字（6763 个）的、以语义类为核心内容的信息库。本文论述了建

立汉字义类信息库的基本思想、收字和立条原则、信息库属性信息的确立、信息库的结构及属性描述、由 6763 个汉字衍生出来的 17429 个字条的归类、基于信息库的各种信息统计等六个方面的问题。这对于 Chinese e-Learning 的研究是很有价值的。

《汉语句法分析中的自动别字纠错》介绍微软研制的汉语文本自动别字纠错系统。该系统所纠正的字均为因音同或形似而易混淆的字。纠错在句子分析过程中进行：句子分析器会同时考虑原文中的字及该字混淆集中的字，正确的字为最佳分析结果之结构树中所出现的字。使用该方法所得到的纠错召回率和纠错精确率均大大超出现有的中文校对系统。目前的中文校对系统中纠正别字的方法，大多数是依靠线性的上下文语境，基本上不涉及句法自动分析，也不依靠表示句法关系的二维的结构树。本文介绍的方法很有新意。纠正在文本中出现的别字可以保证文本的质量，然后才谈得上进行 Chinese e-Learning，这是 Chinese e-Learning 的一项基础性的工作。

《自然语言交际中的语码解读和知识匹配》是一篇理论性较强的文章。文章首先总结了人们对自然语言生成与理解机制的基本看法，进而指出，这些看法忽略了自然语言交际过程的两个重要特点：一个是发话人编码过程中语言信息的 “损耗”与受话人的“重建”；另一个是受话人在同步解码过程中超前的知识匹配。文章指出，这两个特点是自然语言交际的根本性特点，它们既是迄今为止自然语言处理研究的瓶颈，也是未来的努力方向与突破口。本文在理论上对于自然语言交际中语码解读和知识匹配的探讨，显示了作者敏锐的观察力。

相信读者必定会从本期的文章中得到新的启迪，在中文信息的计算机处理研究中取的新的成绩。