唐宋诗之计算机辅助深层研究[1]

胡俊峰，俞士汶

（北京大学计算机科学技术系，
北京大学计算语言学研究所 ，北京，100871）

摘要：介绍了北大计算语言学研究所开发的‘唐宋诗计算机辅助研究系统’。该系统以全唐诗（481 万字）和宋代部分名家诗（160 万字）组成的语料库为基础，运用计算语言学方法对唐宋诗进行分析研究，提取了唐宋诗中的词汇，计 5 万余条目。在对诗文进行词语切分的基础上，建立了词汇的共现关系、对仗关系以及词汇的作者分布特征信息。系统除了提供面向诗文内容的全文检索功能外，还进一步开发了基于词汇的统计分析和诗句相似性检索等功能，实现了对全唐诗的自动注音。

关键词： 语料库语言学，未登录词发现，自动注音，唐宋诗辅助研究
中图分类号： H087,I207

1 缘起

信息技术已渗透到了社会生活的方方面面。在古籍整理方面，基于全文检索的数字化典籍的涌现极大地方便了人们对古籍的整理、分析和研究。已经问世的比较有代表性的数字化典籍系统可以举出台湾中研院的‘翰典全文检索系统’、北大中文系的‘全唐诗电子检索系统’以及台湾元智大学和北大计算语言学研究所合作开发的‘宋代名家诗网络检索系统’等。

随着计算语言学研究的进展，计算机对文本的处理已由传统的录入、存储、全文检索等工作逐渐向文本自动分类、智能检索、信息提取乃至机器翻译等深层次的研究发展。相关技术的进步，特别是在现代汉语信息处理领域的进展为开展古籍的深层次计算机辅助研究提供了思路、创造了条件。

中国古代诗歌源远流长，作为最接近口语的大众化文学形式，诗歌在汉语语言文化中占有极其重要的地位。从唐宋诗入手开展中华典籍计算机辅助研究，无论是从研究语言、文学的角度来看，还是从传承普及中华文化的角度来看，都是一个明智的选择。

北大计算语言学研究所比较早地注意到了这一新的文理结合的研究方向。受 1993 年在北京举行的‘海峡两岸中国古籍整理出版现代化技术研讨会’的启发，计算语言所与北大古文献研究所合作开展了‘古诗研究的计算机支持环境’的研究。1996 年北大计算语言所与台湾元智大学合作，开发‘宋代名家诗网络检索系统’。在此期间，曾应中共中央文献研究室的要求检索了中央领导人在讲话中引用的两句诗‘利欲驱人万火牛，江湖浪迹一沙鸥’（陆游）的原诗。1998 年至 1999 年又承担了国家社科基金‘古诗计算机辅助研究系统及其应用’。

2 系统的开发环境与语料规模

系统在 Windows98 环境下开发，使用 GBK 汉字编码系统。系统选择在中国古代诗歌中最具代表性的‘全唐诗’和宋代部分名家诗为研究对象。录入整理了全唐诗 481 万字，宋诗（包括苏轼、陆游等名家）160 万字。系统采用 ACCESS 数据库格式，将标题、作者、诗文以及派生的辞典信息与知识库信息分别存储在不同的表中，为进一步的深加工创造了条件。

[1] 1998-1999 国家社会科学基金项目（项目号：98BYY022）和北京大学 985 计划项目

3 语言知识库的建立

语言知识库是自然语言处理系统的基本组成部分。面向现代汉语信息处理，北大计算语言所已建立了许多语言知识库,包括各类语法、语义词典以及句法规则库[1]等。但在古代诗歌的研究领域，还没有可以直接利用的成果。因此，建立针对古代诗歌处理用的语言知识库就成为系统开发首先要解决的问题。

3.1 唐宋诗之词汇知识库的建立

系统根据郭锡良教授主编的《汉字古音手册》和广韵资料，录入整理了包含一万多条记录的汉字音韵字典。对其每一个汉字标注了语拼音以及平水韵、广韵等多项信息，为系统进行自动注音、韵律研究提供了必要的基础。

根据观察，多字词在唐宋诗里已经出现了，除了许多双声叠韵词、专有名词外，各类并列、偏正结构的多字词，如：寶劍、北雁、悲傷、安排等，已经被大量使用了。另外，一些词汇，如：白雲、秋風等，虽然在一般意义上可以认为是词组，但由于其在唐宋诗中所具有的特定的隐喻意而使之具有了词的性质。如何正确地识别这些多字词对诗歌的深层次分析有着十分重要的意义。

系统根据汉语的特点,设计了不同于传统互信息方法的一个多维度统计模型,对全唐诗、宋诗语料中的词汇进行自动发现[2]，提取多字词 4 万多条，并完成了人工校对。在参照和利用了北大计算语言所编纂的现代汉语语素库的标注信息的基础上,按照古诗词研究的特点制订了‘ 古代诗歌词语标注规范 ’，并对词典中的 4 万多条多字词和 7 千余单字词完成了词的属性的人工标注。提取唐宋诗中的人名信息 1132 条、地名信息 1392 条。这些词汇信息为实现唐宋诗的词汇自动切分与词性标注创造了条件。

3.2 自动注音知识库的建立

自动注音是指计算机自动对诗歌里的文字进行拼音标注。虽然汉字的现代读音与古音相比已经有了很大的变化，但在大多数情况下，汉字的现代音与古音是一一对应的，确定了一个字的现代读音，也就确定了这个字在古代诗歌里的韵和反切。实现唐宋诗的自动注音不但可以为一般的阅读者提供方便，而且为音韵学的研究提供了第一手材料。

实现自动注音的一般方法是通过上下文信息来确定多音字的读音。系统利用已有的一个注音软件对 160 万字的宋诗语料进行自动注音[3] ，并对注音结果进行人工校对。在此基础上，进行了多音字注音知识的统计提取，提取上下文相关注音知识 14 万余条，并在实际运行中人工修改、添加了 600 多条注音知识，在此基础上开发的自动注音功能对全唐诗的自动注音正确率达到了 99%以上。

3.3 作者信息库的整理录入

在已有资料的基础上，系统尽可能的对全唐诗中的 2821 位作者和宋诗的 5 位作者的背景信息进行了整理，对诸如出生年月、籍贯、性别、类别（分为僧、道、俗、帝 4 类）等信息进行了标注。为作品、词汇以及作者的分类相关研究和时代分布研究创造了条件。

3.4 语料的深加工与相关统计知识库的建立

唐宋诗歌属于历史文献，因此系统设计为一个静态的语言处理系统。也就是说，在保证系统的可扩展性的前提下，系统原则上不考虑诗文本身的动态数据增改。因此，系统可以对所研究的对象预先进行深层次的加工处理以及统计分析。

系统在对所有 600 多万字诗歌语料进行词汇切分的基础上，建立了基于词的全文索引。统计提取了词汇的句内共现关系、对偶位置上诗句之间的词汇对仗关系以及作者的词汇引用关系。

上述所有知识库构成了系统开发的数据基础。系统的总体数据加工流程如图 1 所示。

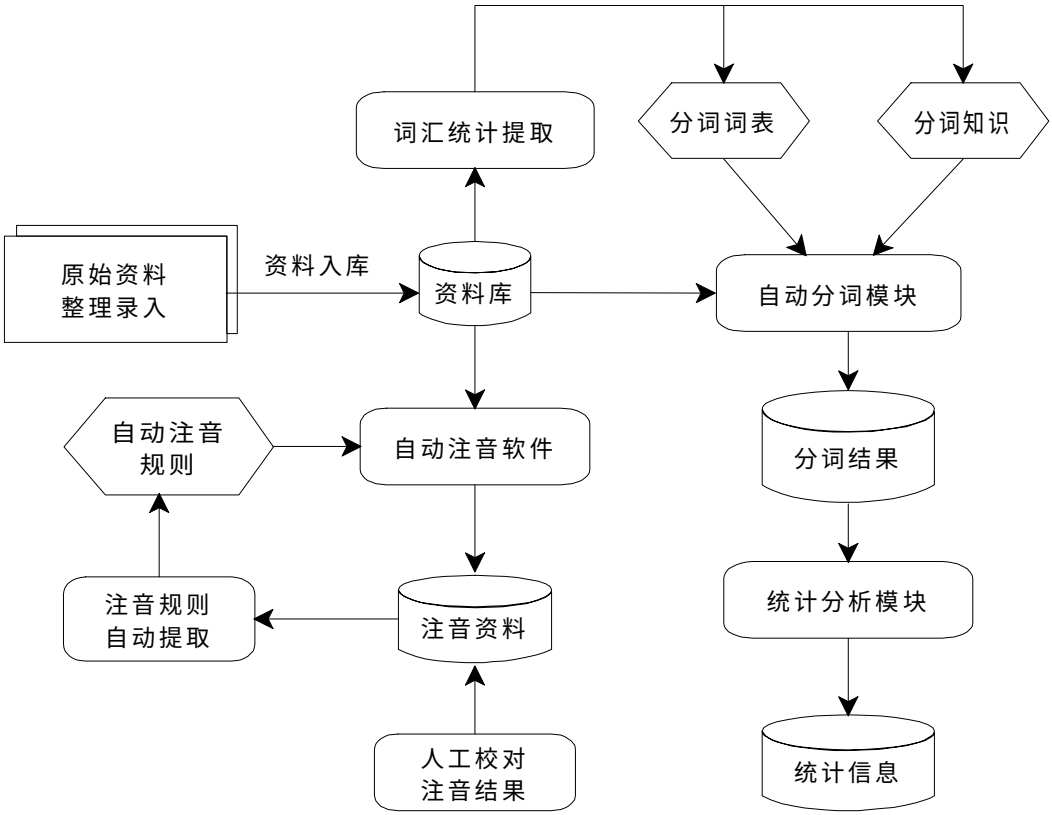

图 1 系统的总体数据加工流程图

Fig.1 Data flow chart of the system

4 系统功能简介

4.1 检索及注音功能

系统对收录的诗实现了按字串、按词的多条件组合检索以及模糊检索，浏览检索结果时用户可以随时使用自动注音的功能对显示内容进行拼音标注。系统提供的基于词汇的诗行、诗句相似性检索为诗句的化用分析提供了线索。

例如盛唐诗人崔颢（704？-754）的著名律诗‘黄鹤楼’在全唐诗中收录有两种版本，主要区别在于首句到底是‘乘白云’还是‘乘黄鹤’，摘前 4 句如下：

昔人已乘白雲（黃鶴）去，
此地空餘黃鶴樓。
黃鶴一去不復反，
白雲千載空悠悠。

在以诗句‘昔人已乘白雲去’进行相似检索后发现，比之较晚的中晚唐诗人刘禹锡（772-842）有‘天上忽乘白雲去，世間空有秋風詞’，意境与用法都十分雷同，而比之较早的初唐诗人张若虚（660？-720？）的‘白雲一片去悠悠，青楓浦上不勝愁’也与本诗的意境十分相似（见图 2）。在对‘昔人已乘黃鶴去’进行相似检索后则没有明显的化用线索。[1]

[1] 刘长卿（？—790？）的诗‘歸沛縣道中晚泊留侯城’中：‘訪古此城下，子房安在哉。白雲去不反，危堞空崔嵬。’与该诗的意境、用法也比较相近。另外还有岑参、寒山等人的诗中也有相似的用法。李白的‘登金陵凤凰台’和‘鸚鵡洲’虽然被认为是‘摹学’，但由于相同词汇不多，无法被系统识别为相似句。另外利用系统提供的模糊检索功能检索到‘乘雲’（包括乘白雲、乘彩雲等）在唐诗中凡 29 见，较之‘乘鶴’（含乘白鶴等）凡 10 见要更加常用一些。

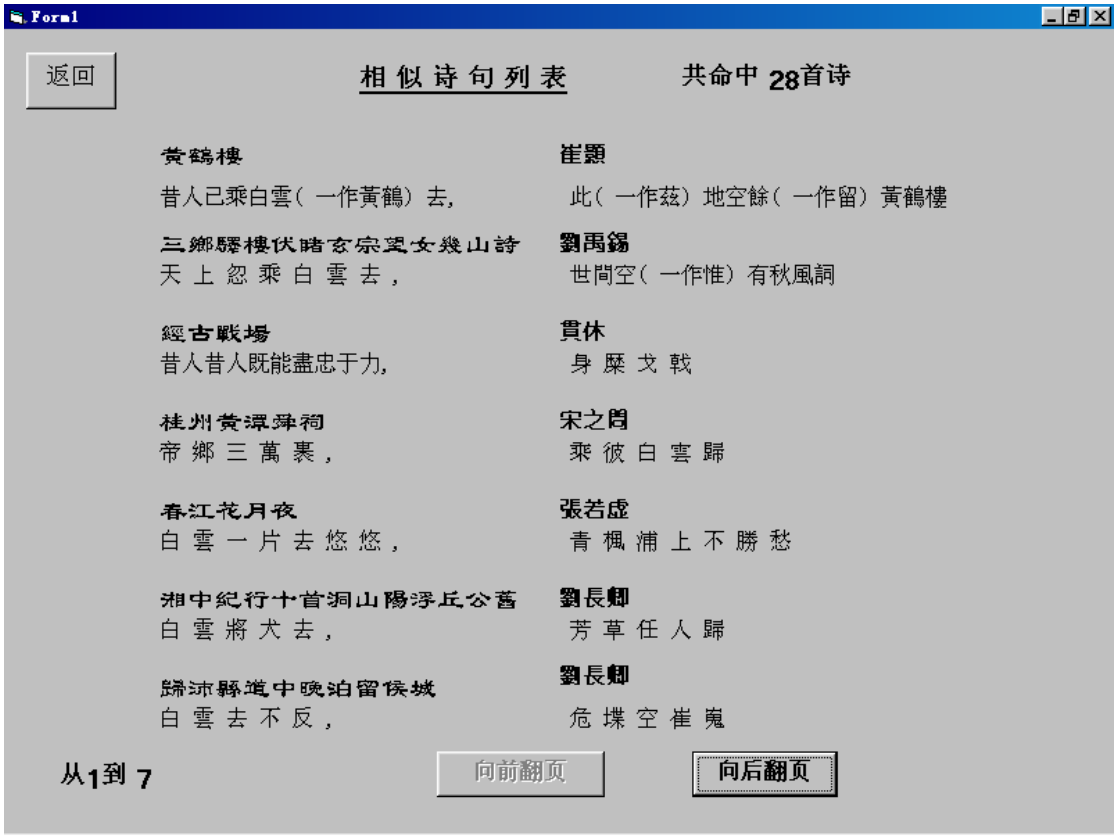

图 2 诗行‘昔人已乘黄鹤去’的部分相似性检索结果

Fig2 The example of the sentence similarity retrieving

4.2 统计分析功能

统计分析是进行唐宋诗辅助研究的重要手段。通过统计可以在大范围内对所研究的对象进行考察，得到许多通过传统的阅读分析很难或根本无法取得的成果。

基于词汇的统计分析是系统的主要特点之一。通过预处理建立起来的词汇之间的共现、对仗以及词汇作者引用关系，提供使用者从多个不同的角度对一些问题进行研究。

例如在考察‘煮’这个动词的共现词汇时（见图 3）就会发现在唐代人们可以煮茶、煮盐、煮饭，但却没有‘煮酒’，‘煮酒’是在宋代才开始有的，因此在三国时代就‘煮酒论英雄’应该是不太可能的[2]。

词汇的使用往往可以反映出作者的创作风格。但仅仅通过词汇引用的次数往往并不能直观地反映作者对词汇的使用偏好。为此系统采用了频度和相对共现度两种标准来体现作者对词汇的使用情况。对于一个词的相对共现度的定义如下：

$$\text{相对共现度 } R = \frac{\text{该词的引用次数}/\text{特定作者用词总数}}{\text{该词的总词频}} * \log(\text{该词的引用次数})$$

图 4 显示了按相对共现度排序的三位唐代著名诗人的多字词使用情况，引用词汇的差别反映了三位作家不同的创作风格。对作品中人名、地名的统计则体现了作者的交游与游历情况。图中最左边一栏是用于参照的全唐诗中相应多字词引用的情况。由于相关的统计是预

[2] 笔者为此参考了北大中文系的‘全唐诗电子检索系统’，对其中收录的唐前诗集进行了检索，也没有发现‘煮酒’的用例。进一步考察了宋诗中的 19 处‘煮酒’用例，发现其中 18 处是用作一个名词词组，如：煮酒不如生酒烈（范正大）；午渴坼瓶嘗煮酒（陸游）。只有一处，煮酒春前臘後蒸（范正大），是动宾结构。可见‘煮酒’在宋代一般是指一种特定的在春前煮制的酒，并非暖酒的意思。该词在辞源中没有收录。

先完成的，用户在查询特定作者的词汇使用情况时不需要进行长时间等待。

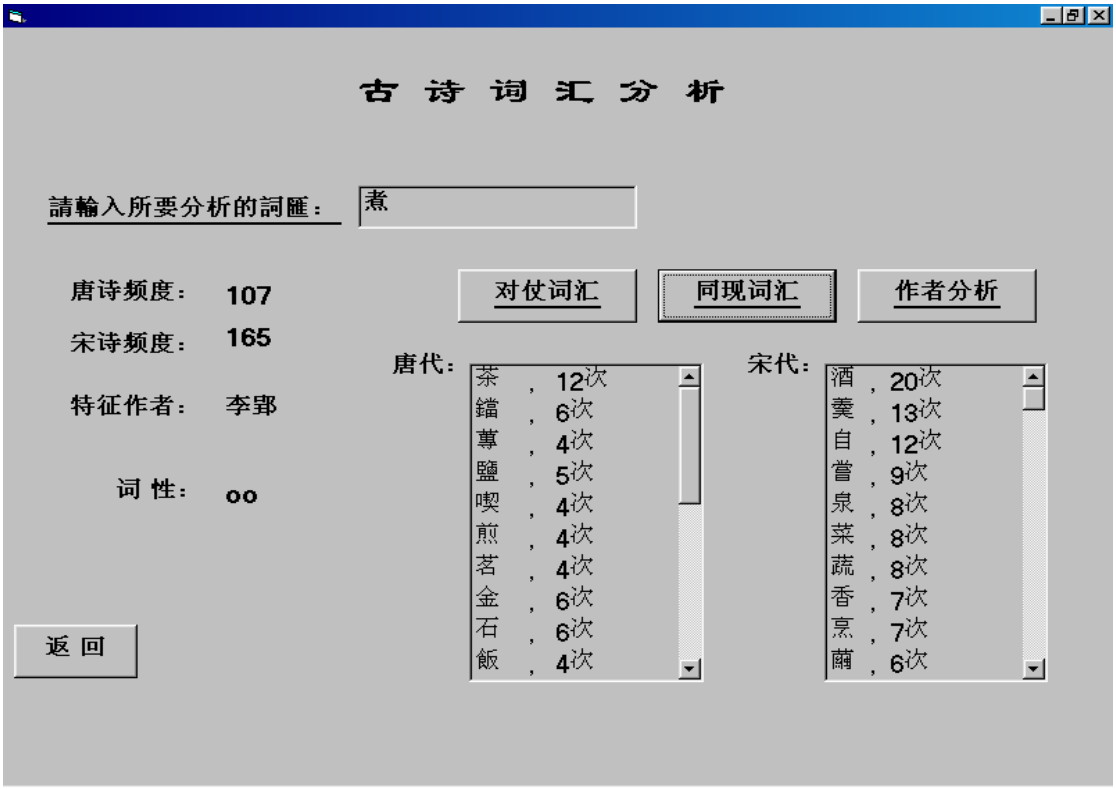

图 3 ‘煮’的同现词汇统计分析结果

Fig3 Part of the collocation of the word 煮/boil

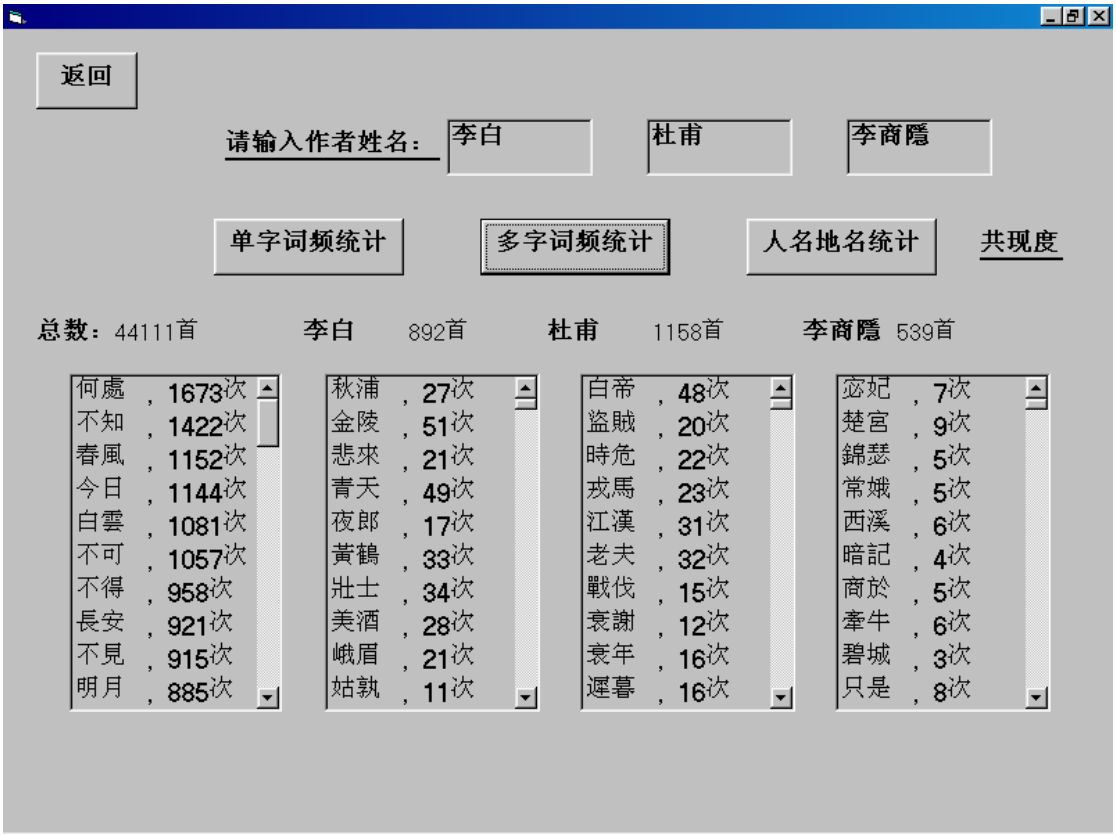

图 4 按共现度排序的作者多字词使用情况统计

Fig4 The multi-char word statistic for different authors

5 一些深层次的研究工作

5.1 意象索引技术的研究

在诗歌语言中常常会大量运用词汇的隐喻意和典故来表达特定的意象。一首描写哀怨的诗可以通篇不见一个‘愁’字。如果需要对具有某个特殊的意象（如思乡）下面的诗进行检索时，单从字串和词入手往往是不够的。鉴于文本的篇章理解技术目前还远未达到实用阶段，因此本项研究的入手点依然选择在词汇一级。在人工选择了与某个特定意象相关的一些特征词汇后（例如：针对‘悲伤’意象选择：悲、苦、愁、凄凉、自憐等），再根据词汇的共现、联想网络搜索到与之相关的词汇（如：蹉跎、蕭然、浮生、西風、殘燈、柳色等共304 条），在此基础上运用神经网络算法对每一首诗的‘悲伤度’进行打分，并据此建立起以‘悲伤’为主题的意象索引。下面两首诗就是意象索引的控制下加入一些限制条件检索出来的表达悲伤情绪但又不含有悲伤词汇的诗。

會昌元年春五絕句　勸夢得酒	答友問
劉禹錫	白居易
誰人功畫麒麟閣，酒客新投魑魅鄉	似玉童顏盡，如霜病鬢新
兩處榮枯君莫問，殘春更醉兩三場	莫驚身頓老，心更老於身

由于缺乏对所收录的诗歌进行人工意象标注的基础，这种意象分类方法的查全率有待进一步验证。就已有实验结果来看，对于‘意象’值较高的短诗（八行以内），实际检索效果较好。

5.2 词汇的时代分布研究

在基于对主要著者的活动年代进行手工标注的前提下，系统对词汇的时代分布特性进行了研究。以 30 年为区段统计了唐代词汇的引用频度的变化情况（对宋诗只区分了南宋、北宋）。在此基础上，可以对指定的词汇群落的时代变迁情况进行分析。由于没有标注每一首诗的具体创作年代，所以在进行词汇引用的时代分布统计时只能通过作者的生卒年进行估算，统计结果的最终意义尚有待于进一步的检验。

图 5 显示了用来表达愁绪的词汇群落（愁、苦、恨、悲、哀）在所研究的语料中按年代的相对引用频度的变迁情况。

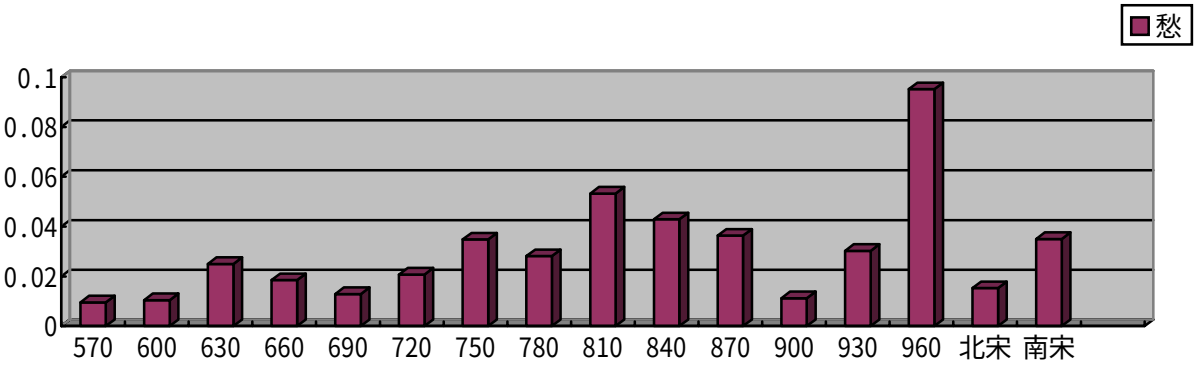

图 5　忧愁词汇群落（愁、苦、恨、悲、哀、忧）的时代变迁分布

Fig 5 The distribution of the word set, sorrow, hate, sad, etc. in different ages

6 结语

运用计算语言学手段对中国古诗词进行研究是一个新的领域，完成的工作大都属于基础性工作，实现的系统也只是一个原型系统。通过本项研究，积累了有关唐宋诗的语料及语言信息知识库，为今后进一步的研究提供了方便。研究所采用的思路多数来源于计算语言学相关领域的研究。反之，本项研究中采用的一些方法，如：基于相对共现度、插入率以及词频的多维度未登录词发现技术等，也可以为现代汉语中有关问题（如半固定短语）的研究提供有益的帮助。

总体而言，对古诗词的分析加工目前还只限于词汇与词汇共现一级，一些相关的应用如：自动切词，相似句检索技术等都是建立在这个基础上的。但作为语言本身还有更高层次的结构（如：句法结构，篇章结构等）。仅在词汇一级进行分析显然是不够的，其相关应用的效果自然会受到局限。可以预见，在这一领域的研究工作还会有很长的路要走。

项目研制得到了台湾元智大学罗凤珠老师的支持与帮助。北大中文系周先慎老师、张鸣老师对本项目的研制提供了许多宝贵的意见。烟台师范大学亢世勇等老师承担了词语标注规范的制定工作并对统计提取的词表进行了校对与标注。北大中文系隋慧娟、徐宝余等同学在作者信息整理和意象分类标准等许多方面为项目提供了帮助，特此表示衷心的感谢！

参考文献

[1]Hu Junfeng,Yu Shiwen
The Multi-layer Language Knowledge Base of Chine NLP, Second International Conference on Language Resources and Evaluation,2000,335:340

[2]胡俊峰，俞士汶
唐宋诗之词汇自动分析及应用，台湾中研院第三届汉学会议

[3] 穗志方，俞士汶，罗凤珠。
宋代名家诗自动注音研究及系统实现。中文信息学报，1998，2：44-53

[4] 罗凤珠，李元萍，曹伟政。
古诗词研究的计算机支持环境的实现。中文信息学报，1997，1: 27-36

The Computer Aided Research Work of Chinese Ancient Poems

HU Junfeng, YU Shiwen
(Institute of Computational Linguistic, Peking University
Department of Computer Science, Peking University, Beijing,100871)

Abstract Based on 6.4 million chars of Chinese ancient poetry, The ‘Computer aided research system of Chinese ancient poems’ provides a word-based analysis platform of Chinese ancient poems. More than 50,000 Chinese words, including 40814 multi-char words, were extracted from the corpora via statistic method. Besides the full text retrieving function, the system also provide word-based statistic analysis, sentence based similarity retrieving, automatic Pinyin tagging and some other useful functions to benefit the profound analysis of the Chinese ancient poems. The National Social Science Foundation of China 1998-1999 funded the project.

Key words corpus linguistic, unlisted word discovery, automatic pinyin tagging, computer-aided analysis of Chinese ancient poems