关于 “契丹小字数字化平台”

吴英喆[1]，白双成[2]

（1.内蒙古大学 蒙古学学院，内蒙古 呼和浩特 010020 ； 2.内蒙古蒙科立软件公司，内蒙古 呼和浩特 010020）

摘要： 契丹小字研究的不断深入，原始资料的陆续出土，对今后的研究工作提出了新的要求。开发“契丹小字数字化平台”、对研究成果的出版、研究手段的改进以及加快契丹小字解读进程均有一定的理论意义和现实意义。该平台由通用契丹小字编辑排版与综合查询系统等两个部分组成。

关键词： 契丹小字；编辑排版软件；综合查询系统

中图分类号码： H211.5　　**文献标识码：** A

一、引言

契丹小字是十世纪初为记录契丹语而创制的文字。随着辽朝的灭亡，契丹文字逐渐变成了死文字。由于它记录许多宝贵的语言信息和历史事实，吸引了语言学家和历史学家以及考古学家的极大兴趣。经过国内外学者近半个多世纪的苦苦探索，到了上世纪七十年代末，契丹小字研究取得了突破性进展。随着契丹小字原始资料的日益增多，研究水平的不断提高，研究工作也遇到了新的难题。一是研究成果的出版常常受到文字的限制。信息技术的迅速发展，印刷事业告别了铅与火，现在普遍采用电脑来编辑排版，过去用扫描、手写等方法处理契丹小字远远不能满足出版业的需求。人们普遍使用 Word 等文字处理软件写文章，如何在文字处理软件中实现契丹小字逐渐成为契丹小字研究者们关注的重要问题。二是研究手段的改善问题。近年来，发现了许多契丹小字原始资料，这无疑对契丹小字的进一步深入研究提供了良好的资料基础，同时对传统研究方法提出了新的要求。古文字研究离不开查找、统计、分类等方法，由于已发现的契丹小字资料比较丰富，用手工进行统计、查找不仅消耗研究人员的大量时间和精力，而且容易产生误差。在电脑技术突飞猛进的今天，我们应该把这样仅靠人的大脑不能完成或者难以完成的繁重的工作让电脑去完成。从电脑中获得的信息，用我们的智慧进行再分析整理，然后把分析的结果再输入电脑进行有效地加工。这样把电脑作为辅助工具，为我们的研究工作提供有效的帮助，尤其对契丹小字这样死文字，只有人机结合的研究方法才能尽快揭开其中的奥秘。据说，早在上世纪 70 年代，苏联学者斯达里科夫等人利用统计学的方法对契丹小字组字、语法规律进行过种种探索，由于当时所发现的契丹小字资料过于稀少，他们未能达到预期目标。可是现在，据不完全统计，已发现契丹小字的总数已达 3 万字。因此，利用计算机进行研究的条件已成熟。由于契丹小字笔画繁多，构造复杂，字形整理工作还没有完成，而且还需要投入大量的人力、物力、财力，使得契丹小字字库建设迟迟未能实现。过去我们委托软件公司制作过契丹小字字库，但由于种种原因未能推广使用。

为了解决以上问题，根据内蒙古大学蒙古学研究中心承担的教育部人文社会科学重点研究基地重大项目“契丹小字再研究”任务书，于今年 9 月，“契丹小字再研究”课题组与内蒙古蒙科立软件有限责任公司达成协议，签订了“契丹小字数字化平台合作开发协议书”， 决定联合开发通用契丹小字编辑排版与综合查询系统，统称为“契丹小字数字化平台”。笔者认为，该平台一定能够满足研究、出版单位以及从事契丹小字研究人员的需求。下面简要介绍该平台的有关情况。

二、契丹小字编辑排版软件

契丹小字编辑排版软件是在 Windows 环境下，通过 MS office 的二次开发接口，扩展 Word 来实现。原则上保留 Word 原有的功能的基础上，增加契丹小字编辑排版功能。具有安装方便、操作简

单、功能强大等优点，实现真正意义上的“所见即所得”。鉴于契丹小字是一种表音文字，我们对其采取的字体编码方案是对原字进行编码，而对原字的组合框架在编辑层次中处理。这样做的好处是可以将新发现的原字随时加入字库中，不会对原字组合产生任何影响。编码方案上为了达到多语种的混排，并考虑到契丹小字本身没有国际标准编码，暂且放在 Unicode 自定区中。

（一）字库开发

契丹原字字库是选用频率高、具有代表性契丹原字字形构成的字库。该字库中适当收入异体字，但不收入破体字，并留有空码位，以便及时补充制作新原字。它要求选择的字形具有代表性，一般以《宋魏国妃墓志》和《太叔祖哀册》中出现的字形相对比较规整的原字为蓝本，尽量向《契丹小字研究》[1]所规范的原字形体靠拢。契丹小字原字中存在着许多异体字，如何从众多的异体字中选出真正的通用原字是契丹小字研究者们一直关心的问题。辨别异体字也是契丹小字研究的重要内容之一。但作为编辑排版软件我们尽量把异体字收入进去，以便文章中展开讨论。字库中专门制作方框和空位，以备研究人员描画破体字的雏形。传世的契丹小字多数为用刀或其它坚硬的工具刻在石头上的文字，不同于毛笔或钢笔书写的风格。所以制作其字模应当表现其原有的风格。已出土的契丹小字散失到各地，且今后还要陆续出土，所以不便确定其原字总数，需要一边收集整理一边制作其字模。

（二）原字输入法开发

输入法中包括两个层次，首先是原字的输入，其次是小字的输入。鉴于多数原字的读音尚不明确，契丹原字的输入采用内码输入和字型输入。

1．内码输入法：实为编码输入，为简便可改为序号输入，具体的编号可以有多种

方案。但不管用什么方案都不能成为批量输入手段，所以就不展开讨论。

2．字型输入法：有两种实现方案，一是只对第一笔画（或首二笔画）分类，其余

候选。二是对所有笔画进行编码。绑定键，类似于五笔输入，这样可以提高输入速度。契丹小字的 500 多个原字，按其第一笔画（或首二笔画）的特征可分为 ：一类原字、十类原字、𠂇类原字、𠂆类原字、丿类原字、乙类原字、亻类原字、八 类原字、𠂉类原字、丨类原字、冂类原字、 丶 类原字、ヲ类原字等 13 个类型。每一类原字的键盘上的代码分别如下：

一类为 H 、十类为 J 、𠂇类为 N、𠂆 类为 F、丿类为 P、乙类为 E、亻类为 I、八 类为 B、𠂉类为 K、丨类为 S、冂类为 M、丶类为 D、ヲ类为 W。

输入其代码，契丹原字以 10 个为一组显示在屏幕上，然后选择其对应的候选字，完成输入目的。如果第一组中没有显示要输入的原字，继续按+键翻页，然后选择对应的数字，完成输入目的，其余类推。每个类型中的常用原字一般居于第一组，而且每输入新的原字，将改变原字的位置，新输入的原字自动显示在第一个位置。

此外，菜单栏中设置一个契丹小字原字总表，以便从字库中直接点击的方式录入契丹原字。

（三）编辑排版

目前 Word 中文版已成为我国最为流行的文字处理软件之一。为了契丹小字字处理的特殊要求，该软件除了保持中文版 Word-XP 的所有功能，还新增了以下几种特殊功能：

1.多种文字混排功能：该文字处理软件中还能实现 Word-XP 原来不能支持的中国北方民族文字。如：蒙古文、八思巴文、回鹘式蒙古文、阿里嘎里字母、索永布文字、托忒蒙文、满文、锡伯文等。

2．输入国际音标：契丹小字研究者们经常使用 Word 中不能实现的国际音标，为了满足这一特殊需要，契丹小字文字处理软件中建立支持比较常用的国际音标的输入排版功能。

3．竖排及注旁译功能：传世的契丹小字原始资料均为由上而下书写，从右到左换行，契丹小字

文字处理软件不但能够实现这一书写规则，而且行间可以简便地注明已释读词语。

4．表格中插入契丹小字的功能：契丹小字研究论著中经常需要插入表格，该软件中可以随意地将契丹小字插入表格，并进行各种格式设置。

5．处理不同结构字的功能：契丹小字合成字的书写规则一般为从左到右，二二下推，如果是 3、5、7 个原字组成的小字，其最后一个原字居于中间。我们把这样的常用结构设为默认，即原字与原字间不写任何符号或空格，最后按回车键，屏幕上自动显示其常用组合形式。根据后来出土的资料，我们发现契丹小字的书写规则也有特殊情况的。即以 2、3、4、5、6、7 个原字组成的小字也有首一原字居于中间的。契丹小字文字处理软件对这种特殊结构字的处理方法，采取了最后一个原字前写逗号的方法，满足特殊结构词的输入目的。

6．文字修饰功能：用户可以像处理其它普通文字一样，对契丹小字进行随意的复制粘贴及各种修饰，包括字体大小，斜体，粗体，阴影，颜色等。

7．系统提供了方便的契丹小字修改功能。用户可以通过编辑框对文档中的契丹小字进行替换、删除、添加，原字的组合也随之进行自动变换。

8．表示模糊字、破体字、残体字的功能：出土的契丹小字文献中经常遇到模糊、破体、残体字。系统中专门设置了原字外围打虚线方框的功能，以表示模糊字。字库中还制作了空位和方框，可以与其它原字组合，以便表示破体字和模糊字。

三、契丹小字综合查询系统

我们把契丹原字输入法开发和字库建设作为基础，打算创建一个数字化的契丹小字综合查询系统。目的在于为契丹小字研究提供一个现代化的手段，提高研究效率，加快契丹小字解读进程。对于契丹小字，虽然它有了较长的研究历史，并且达到了一定的研究水平，然而过去的研究方法一般都着重于比较研究，利用统计学的方法或其它新的手段来研究契丹小字的却不多。近年来，虽然发现了许多契丹小字墓志哀册，但由于研究手段和研究工具方面没有进行明显的改善，一定程度上影响了契丹小字的解读进程。据于以上思考，我们认为只有综合传统的自然语言处理技术和基于统计的自然语言处理技术才能提出更有效的解决方案。我们的研究方法和思路是数据驱动的方法：收集大量契丹小字原始资料和相关语料以及研究论著，从中获取语言知识和破译知识。为了支持知识获取工作效率的提高，我们将建立大规模的契丹小字研究图库、契丹小字语料库、汉字文献语料库、研究资料库等，并对数据库进行不同深度的标注、关联。下面简要介绍该系统的基本框架：

（一）数据库

1． 契丹小字研究图库：包括契丹小字墓志哀册的拓片、汉字墓志哀册拓片及其它。由于契丹小字长年久月地深埋地下，加之出土时的损伤，造成很多契丹小字有明显的裂痕、残缺，以此为原型的契丹小字拓片有许多模糊难辩之处。我们打算利用扫描和反拍等方法把有关拓片输入计算机，用图形图象软件进行处理，添加墓志名称，写行数，以便检索查询。

2． 契丹小字语料库：包括迄今发现的契丹小字金石资料。这是最关键的语料库，整个数据库的核心，所以我们首先重点建设该语料库。过去我们建立过类似这样的语料库，但由于种种原因未能满足我们的要求。这次我们在编辑排版中解决了模糊字、破体字、残体字的处理方法，而且还制定了随时添加新原字的方案，所以大大减少了建设该语料库的难度。为了检索需要我们对数据库中的已释读词语做适当的词性标注。

3． 汉字文献语料库：包括汉字石刻及史籍中的与契丹小字解读有关的历史资料。如：契丹语、人名、地名、官职名、人物传等。该语料库的建设需要大量的人力和物力，而且可能需要相当长的时间。除了校对，摆在我们面前的难题是如何处理那些繁体汉字的问题。虽然 Word 的中文字库中收入了非常丰富繁体汉字，但还不能很好地满足我们的要求。

4. 研究资料库：包括 1977 年以后发表的契丹小字解读论著的摘要。自从上世纪二、三十年代到今天，有关契丹小字研究论文已发表三、四百篇。1977 年以后契丹小字研究进入新的历史时期，因此，这个资料库中收入的论著以 1977 年为界线。而且与解读无关的一般的介绍性论著，暂且不准备收入。内容包括：契丹小字、释义、拟音、关键词、研究人员、论著名称、发表年月、刊物或出版社名称、起止页。这个数据库能够对扩大释读范围，检验解读结果以及学习研究者提供很多方便。

（二）、检索

1．图片检索：输入关键词的方式能迅速查找图库中的任何一张契丹小字或汉字文献拓片。根据需要，也可以检索拓片的某一部分，并且以放大、缩小等方法能清楚地浏览图片。

2．汉字检索：输入关键词能够迅速、准确地从数据库中获得有关信息。

3．契丹小字检索：包括字、词、附加成分等不同角度的检索功能：

（1）原字检索：查找某个原字在文献中所处的位置。查询结果包括文献名称、行数、该行中的第几个字、合成字的第几个原字、出现次数、出现频率。

（2）单体字检索：查找单体字在文献中所处的位置。查询结果包括文献名称、行数该行中的第几个字、出现次数、出现频率。

（3）合成字检索：查找合成字在文献中所处的位置。查询结果包括文献名称、行数、该行中的第几个字、出现次数、出现频率。

（4）排列：以指定的格式，指定的顺序排列语料库中的所有契丹小字。排列结果包括文献名称、行数、该行中的第几个字、出现次数。

（5）原字出现情况：检索某个文献中的原字出现情况。检索结果包括文中出现了哪些原字、未出现的有哪些原字、每个原字的出现频率、出现次数。

（6）单体字或合成字出现环境：检索指定单体字或合成字的前后若干字。检索结果包括：文献名称、行数、出现次数、出现频率。

（7）静词索引：检索名词（人名、地名、官职名）、数词、代词、形容词。检索结果包括文献名称、行数、该行中的第几个字、出现次数、出现频率。

（8）词索引：检索结果包括文献名称、行数、该行中的第几个字、出现次数、出现频率。

（9）词与词的关联：首先确定一个词语，以它为中心，可以检索其前后，是否出现与它有关的词语。检索结果包括文献名称、行数、出现次数、出现频率。

（10）附加成分索引：包括静词附加成分、动词附加成分。检索结果包括文献名称、行数、该行中的第几个字、出现次数、出现频率。

（11）附加成分与附加成分的关联：先确定某个附加成分，以此为主体，能够查看其前后出现的有关附加成分。检索结果包括文献名称、行数、出现次数、出现频率。

完成契丹小字数字化平台的开发任务之后，打算建立契丹大字和女真文字字库，逐步把平台建设成“契丹、女真文数字化平台”，条件成熟之后，我们打算利用 Internet 来管理平台中的各项数据。

目前，契丹小字数字化平台的开发工作正在顺利进行。现阶段，平台的各项功能还不完善，还

没有达到我们预定的目标。到目前为止，契丹小字字库的雏形已经确定，下一步要大量的测试和修改字模。预计今年 11 月 15 前编辑排版软件问世，12 月 31 前检索系统问世。为了让读者亲眼目睹我们的成果，特意把尚未定型的契丹小字及其编辑排版例样附于文后，并期待更多的专家和学者对这一项目给予关注。

参考文献

[1] 清格尔泰，刘凤翥，陈乃雄，于宝林，邢复礼.契丹小字研究[M].北京:中国社会科学出版社,1985.

Digital Platform of Khidan small script

WU Ying-zhe[1]，BAI Shuang -cheng[2]

（Mongolian Language Institute, College of Mongolian studies,Inner Mongolia University,Huhhot，010021 China）

Abstract: The further studies and unearthing of firsthand materials of khidan small script put forward some new challenges to our future study. Developing the digital platform of khidan small script is of some theoretical and current significance on such three aspects as publishing the research works、renewing the means of research and accelerating the process of comprehending khidan small script. The digital platform is consist of editing 、typesetting and synthetic searching system of khidan small script .

Key word: khidan small script ;editing and typesetting software ;synthetic searching system

收稿日期: 2004-10-21;

基金项目: 教育部人文社会科学重点研究基地重大项目（02JAZHJD840007）;

作者简介: 1、吴英喆（1971—），男，蒙古族，内蒙古通辽市人，内蒙古大学蒙古学学院，讲师，博士研究生;2、白双成（1974—），男，蒙古族，内蒙古乌兰浩特人，内蒙古蒙科立软件有限责任公司副总裁。