字与词的义项分析

邹晓辉、邹顺鹏
北京市海淀区学院路 29 号 100083 qhkjy@yahoo.com.cn
中国地质大学（北京）思想政治教育学院高等教育研究所

摘 要

本文回顾了汉语研究的几次“本位”转换，认为这就是前辈在探寻汉语理论体系建构的逻辑起点。如果必须在字与词之间做出选择，那么，赵元任“字是中国人心目中的中心主题”的论断，以及徐通锵“‘词’不能代表‘字’而成为汉语的一种基本结构单位”的论断，都选择了字。而之前，其他人分别选择了词（马建忠 1898）、句（黎锦熙 1924）、词组（朱德熙 1982）。这是为什么呢？难道赵、徐两位语言学教授都不懂字和词的常识性义项吗？显然不是。他们是根据学术直觉作出上述断言的。有没有可支持他们这种学术直觉背后的学术性义项呢？这是笔者提出的问题和思考。本文从宏观的类与微观的例两个层面多个角度，分析了字与词的学术性义项，发现常识性义项往往是造成非学术性争议的起因。本研究采用史论结合形式和学术分类方法，通过对“语”与“字”两个亚系统的二八分类，凸显了音节、字、言“迭交”的特点，得出了“汉语就是组字成语的典例系统”的论断。陆俭明、冯志伟都说如果字本位换成语素本位，也就不会有人再反对字本位了。因此，为避免字本位提法所引起的非学术之争，与其说字不如说言是汉语的基本结构单位，从而便于我们集中精力论述字素（周上之）、字词（潘文国）、字语（邹晓辉）几组术语及其关系（这是字本位内部尝试统一认识的努力之一）。

关键词：普通语言学、形式信息学、汉语形式化、汉语研究的逻辑起点、字本位、言本位

一、 引言

本文旨在通过探讨《字本位与中文信息处理的基础》（简称《基础》）中涉及汉语独特性的几组关系，即：字与素、字与辞、字与语、言与文，明确指出汉语组字成语的特点，同时，通过分析字与词的范畴迭交，明确指出汉语从古到今的发展演变过程中因外语的影响事实上已是中西两类语言范畴基本框架并行格局，古、现代汉语两个相对独立的体系存在就是例证。

有无可能在普通语言学和形式语言学两个层面来统一汉语理论基础架构呢？

本文由“字”和“语”二八亚类之中遴选出“言”这一基本范畴及术语，希望用它指代汉语基本结构单位，以此明确语言和文字的界限，可避免“字本位”以字称谓汉语基本结构单位遭遇“语言与文字的界限不分”的根本误解。

这样做的好处是可让字本位正反双方的语言学者集中精力探讨汉语自身的几组关系，即：字与素、字与辞、字与语。因为，言与文的区别是显而易见的，提“言本位”不应该再会有“语言与文字的界限不分”的误解。

这样做还有一个好处是可让“字本位”正、反双方的语言学者集中精力探讨现代汉语因外语的词范畴导入必然与现代汉语与古代汉语一脉相承的“字”或“言”范畴之间发生中西两类语言范畴体系“迭交”而产生的一系列问题，如“字素冲突”、“字词冲突”以及“汉外双语在语汇、语义、语法乃至语用习惯等诸方面的冲突”。

可以说汉语研究一百多年在“词本位”（马建忠）、“句本位”（黎锦熙）和“短语本位”（朱德熙）之后重返“小学”传统的“字本位”（徐通锵）的探讨，理应更为深刻而积极。汉语学界无须因为“字->词->句->词组->字”看似回归的表面现象就简单地误解“字本位”立言的深意。百多年汉语学界刻苦艰辛地探索，绝不会只是又简单地回到了“字”。

本文论述框架：（一）字词关系凸显，（二）类的宏观分析，（三）例的微观分析。

二、 正文

本文从一个特定角度概要介绍《基础》有关汉语独特性的探讨。

（一）字词关系凸显

由于“中文信息处理需要的，并不是现在汉语学界已有知识的照搬：有的方面需要根据计算机的‘能力’去总结汉语的规律，在一定程度上，还需要研究者抛开传统语言学的固有习惯和方法；有的方面则需要填补上已有知识的不足。”[1] 因此，本节采用以下框架开题：

“以词称字”——“词本位”（马建忠 1898）

“用句辖词”——“句本位”（黎锦熙 1924）

“词组代句”——“短语本位”（朱德熙 1982,1985）

“字词有别”——“字本位”（徐通锵 1992，1997，2005）

图 1 是汉语研究探寻体系建构的逻辑基点更迭次序“字->词->句->词组->字”示意图。

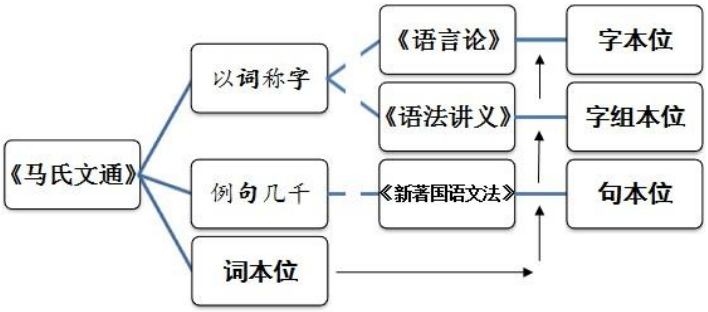

图 1

由图 1 可见，四个箭头指示出的 3+1 个“本位”学说的更迭次序。说 3+1 是因为一方面“字本位”还在探讨之中，暂无定论；另一方面，前三个与后一个“本位”的含义不尽相同。图中《马氏文通》[2]字、词、句，都涵盖了，就是没词组。这是为什么呢？因为它论述的古文例句是依据“因字而生句，积句而成章，积章而成篇。”[3]而生成的。 还有一个问题，《文通》写的是字，为什么汉语学界要说它是词本位的肇始呢？因为它借鉴的是拉丁语以及法语的词本位，即以词法为重点、以词类为基础来描写语法现象。

黎锦熙借鉴英语句本位认为《文通》“仅就九品词类，分别汇集一些法式和例证，弄成九个各不相关的单位，是文法书最不自然的组织，是研究文法最不自然的进程。”[4]

龚千炎曾评价“从总体看，‘句本位’显然要比词本位进步，因为它不是孤立地静止地看问题，而是从整体中看个体、从动态中看语言结构”。

朱德熙借鉴美国结构主义短语本位认为句本位“这种语法体系里，由于词组、句子成分，中心词等基本概念之间，互相不协调，产生了许多矛盾。…缺乏严谨性，同时也缺乏简明性，实在不能说是一个好的语法体系。”并在《语法讲义》里采用了词组本位。[5] 朱德熙认为“由于汉语的句子的构造原则跟词组的构造原则基本一致，我们有可能在词组的基础上来描写句法，建立一种以词组为基点的语法体系”。[6]

人们**发现**，词组本位具有词本位和句本位无法比拟的优点：由于汉语句子的构造原则跟词组的构造原则基本一致，所以以词组为基点描写句法，内部一致，没有矛盾；同时词组的结构讲清楚了，句子的构造也相应的讲清楚了，用不着分两套讲，显得严谨、简明而又自然。

但是，后来进一步研究**发现**，词组本位存在一个基本问题，这就是：因为词组由词组成，而“词”是印欧系语言的基本结构单位。传统的汉语研究只有“字”而没有“词”。而词是《文通》从印欧语中移植进来的，在汉语中没有根基。赵元任（1975 年，国内 1992 年 233-234）认为印欧系语言的 word（词）这一级单位“在汉语里没有确切的对应物”。[7]

于是，新的问题又出来了。试想：如果没有词，又哪来词组呢？从词本位到句本位进而到词组本位，都是“外来的理论”在汉语中没有根基。当初《文通》字词不分的问题出来了。

这的确值得学界反思！字本位（徐通锵）观点的探讨，可视为字词关系在理论上的凸显。

笔者自 2000 年由徐通锵教授领进其以汉语为例而论述普通语言学的理论研究领域，并对“字本位”从借鉴回到传统的顿悟以及它遭遇的几点质疑（字以及核心字等基本概念定义的困难，言文、字词以及字素几对矛盾造成的理论冲突）进行了独立的思考，从汉语形式化的角度，提出了音节总量控制模型（GSCM）和文本总量控制模型（GTCM）以及“层面型结构”与“线串型结构”及其“迭交”的观点[8]。徐老师（作为青岛会议主要审稿人之一）看后认为笔者“从层面型结构与线串型结构的迭交着眼去定义字，这一思路很好”（徐通锵 2005）。

笔者认为，立足于“音字”作为“线串型结构”的“节点”（含“起点”）是“字本位”避免“语言与文字的界限划分不清”误解的一个具有建设性、启发性的观点。特介绍如下：

图 2 是层面型结构与线串型结构“迭交”原理应用于计算机字库设计的例证示意图。

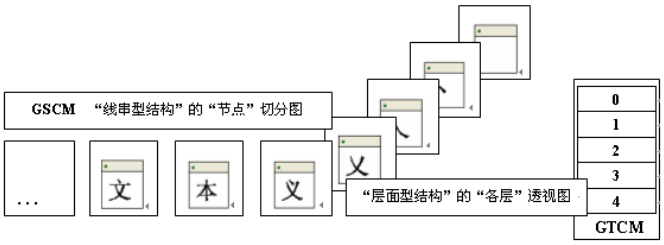

图 2

由图 2 可见“音字”切分为“节点”与“形字”拆分为“部件”。“层面型结构”顶层可透视的音形“迭交”情形。在图 2 中“义”这个“字”正位于“线串型结构”的“音字”（语言可分出库）与“层面型结构”的“形字”（文字可简化入库）的“交汇处”。[9]

2002 年黄昌宁教授邀请笔者到微软亚洲研究院为其介绍字本位与中文信息处理的研究成果（邀请笔者的还有中科院的黄河燕、北大的俞士汶、清华的陈群秀、中软的关维忠等）。2007 年黄昌宁表达了对字词关系的新认识：“把分词过程视为字的标注问题的一个重要优势在于，它能够平衡地看待词表词和未登录词的识别问题。在这种分词技术中，文本中的词表词和未登录词都是用统一的字标注过程来实现的。在学习架构上，既可以不必专门强调词表词信息，也不用专门设计特定的未登录词识别模块。这使得分词系统的设计大大简化。在字标注过程中，所有的字根据预定义的特征进行词位特性的学习，获得一个概率模型。然后，在待分字串上，根据字与字之间的结合紧密程度，得到一个词位的标注结果。最后，根据词位定义直接获得最终的分词结果。总而言之，在这样一个分词过程中，分词成为字重组的简单过程。然而这一简单处理带来的分词结果却是令人满意的。”[10]

（二）类的宏观分析

图 3 是《基础》第一部分目录关系解说（涉及字词两组可形式化的基础“类”）示意图。

1.**语言**（对象语言与解释语言的关系）

1.1.**字**（古）

1.2.**词**（今）**【中】**

1.3.**Word** **【外】**

（自然人&大脑）

2.**信息**（自然语言形式处理殊途同归）**（人机互补双脑协作）**

2.1.**Zi**（古）

2.2.**Ci**（今）**【中】**

2.3.**Word** **【外】**

（计算机&电脑）

图 3

由图 3 可见，笔者强调语言的两“类”形式区分，旨在便于计算机针对这些特定的“类”建构实用的标注序列，即：对象语言“例”义项分析可标注重用的解释语言数据库。

从形式语言的角度看，汉语的符号对象只有一个类，这就是字；字内外的其它类（可由两个系列的各种亚类组成）都是解释字例义项的具体组合形式。同理英语的符号对象也只有一个类，这就是 word（词）。现代汉语的词不过是古代汉语的字与外语 word（词）混合的一个类。这些是对于人脑的基本形式分类；对电脑而言在美国标准信息交换码（ASCII）的字符库里三者都必须化为同一个类，即字母组合形式。

图 4 是汉外两种等价的语法称谓体系(分别基于字和词而建构的两个理论框架)示意图。

▪ 章法 ▪ 篇（起承转合围绕主题） ▪ 段（起承转合围绕段意）	
▪ 特点：辞句同构、辞块等价	▪
▪ 句法 ▪ 句（语意停顿） ▪ 读（语气停顿）	▪ 句法 ▪ 句（单、并、复） ▪ 句子成分
▪ 组字法 ▪ 块：虚实结合 ▪ 链：虚字组合 ▪ 辞：实字组合 ▪ 言：音字（解释语言）	▪ 词法 ▪ 虚实结合的词组或短语 ▪ 虚词组成的词组 ▪ 实词组成的词组 ▪ 词（解释语言）
▪ 造字法 ▪ 字：形字（对象语言） ▪ 偏旁部首：字中字 ▪ 偏旁部首：变形字 ▪ 偏旁部首：缺损字 ▪ 基本笔画	▪ 构词法 ▪ 词素或语素（对象语言） ▪ 词根 ▪ 词缀（前、中、后） ▪ 词头、词尾 ▪ 拼音字母

图 4

由图 4 可见，汉语组字（基本结构单位）成语（言、辞、链、块、读、句）的过程具有“辞块等价、辞句同构”的特点。

如果以一种双语对等的视角来看，那么，语汇一级，汉语的字内外组合变换与英语的词内外组合变换，除了符号（大小字符集）组合形态的差异之外，就是各自标音取意的方式之不同，其中，最突出的区别就是汉语音字的单音节特性和英语单词的混音节特性所造成的彼此自身与后续结构单位（如汉语的辞链块和英语的三类对应的词组）之间如何相互区别的方式方法的不同，正是该区别决定了汉英两种语言形态之间的巨大差异，即：汉语丰富的字内语素及其形态特征表现为偏旁部首的显性特征被固着于形字，音字分合而组成辞、链、块的简单过程，一方面，因“形字”与其“迭交”所含的内部语素（偏旁部首）已固着而几乎已隐性化了，另一方面，由于引入词的范式而带来的不能独立使用的字间语素与可以独立使用的言、辞、链、块之间可能存在的某些相互嵌套关系，致使音字的属性变得复杂；与之相反，英语（形态变化丰富其他的西方语言，如德语、法语和西班牙语等等更不必说）则由于其词形的变化可因其在句中的具体地位不同而改变其相应的具体形态，从而使其内部语素（词内语素）及其形态特征几乎都显性化了。

图 5 是汉语“字”和“语”二八分类直观描述（文字和语言的关系清晰）示意图。

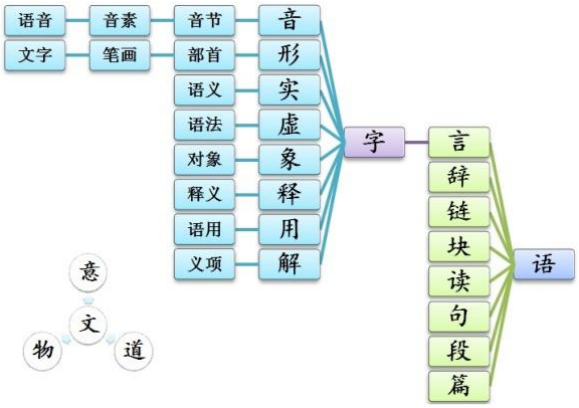

图 5

由图 5 可见，作为汉语基本结构单位的“字”或“言”位于“辞、链、块、读、句”等衍生组合形式或其后续结构单位的逻辑起点。

必须指出，字和语的分类蕴含字词关系，而字词的范畴迭交却蕴含中西双方复杂的背景。

图 6 是汉语“字”和“语”二八分类的进一步详解（可很好地区分语言与文字）示意图。

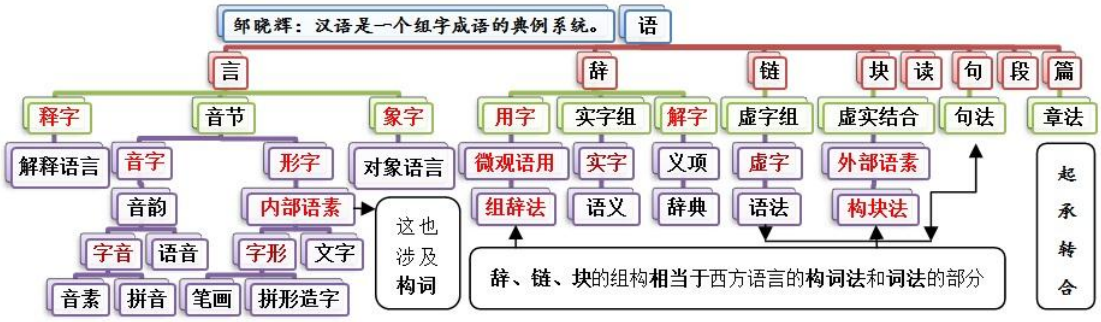

图 6

由图 6 可见，“语（言、辞、链、块、读、句、段、篇）”与“（音、形、实、虚、象、释、用、解）字”之间“迭交”于“言”这一单音节的汉语基本结构单位，汉语的其它结构单位均可视为单音节的“言”的衍生组合形式或后续结构单位。古今的汉语均满足该特点，

在图 6 中，“内部语素”和“外部语素”以及“微观语用”反映字词关系以及字素关系。笔者提出：汉语研究需要把“字内语素”和“字间语素”以及“非语素字（可独立使用的言）”分别以“内部语素说”和“外部语素说”以及“微观语用说”的划分方式进行专门的研究。

进而，可通过“字”或“言”的义项分析，相应地建构《义项字典》和《用例辞典》，与此同时，还可相对独立地汇编《汉语链表》和《语块手册》。

其目的是：一方面，从语义和语法的角度深挖汉语的字内与字间两种语素形态组合变换的形式信息，另一方面，如图 4 显示可以在汉语的辞、链、块和英语的三类对应的词组之间，寻求汉语的字与英语的词之间在义项层次相互对接的纽带。

由此引出：由字向词（即汉译英的进路）和由词向字（即英译汉的进路）的义项分析或双向标注。通过大学课堂的计算机辅助双语教学中的汉语的字与英语的词的义项分析或研究，可把汉英-英汉（地道而常用的言辞或词语）双语转换工具的使用以及验证的过程融入上述义项分析或双向标注的过程，可让在校生有机会参与这种创造性合作型生产式融智教研活动。

（三）微观的例分析

《基础》第二部分，即：微观分析工具。涉及字与词的义项分析（即精准知识处理）的人机协作方案。

图 7 是《基础》第二部分目录（标明了形式信息处理和内容信息处理的关系）示意图。

图 7

由图 7 可见，“计算机辅助双语教学中字与词的义项分析”的方法及结果。其中，“三解”及“三集”是**形式**信息处理；“三注”是**内容**信息处理；“三表”则是形式与内容的结合部。

图 8 是微观分析工具的设计原理（“孪生图灵机”）示意图。

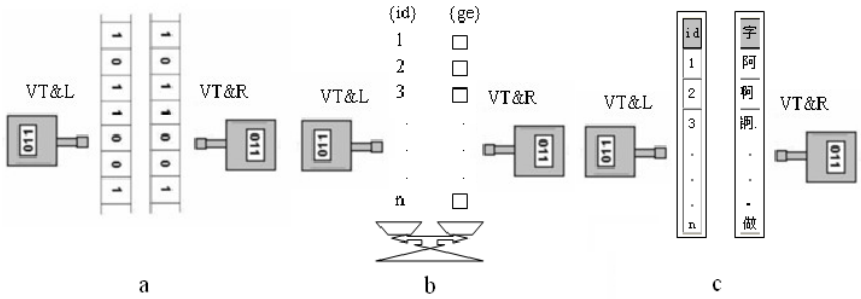

图 8

由图 8 可见，左边孪生图灵机 a（与中间的 b 和右边的 c 均等价），其特征在于 b 所述的天平式计量转换装置是基于“同意并列、对应转换”法则而构造的，其具体使用方式由 c 基于可穷举汉字集而构建，通过标准化与个性化结合的“双列表”实现数-字分工协同计算。

图 9 是体现人机协作互补的一整套约束机制（具有“冯氏多胞机”存-处功能）示意图。

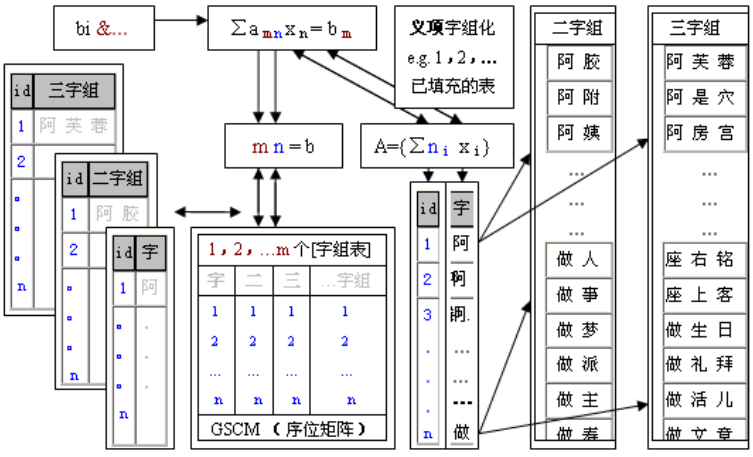

图 9

由图 9 可见，一系列的“双列表”【即“孪生图灵机”的一种表形式（即“多胞冯氏机”）】，左列序号“音节总量控制模型（GSCM）”数据自动查询经“双列表”自动转换可有针对性地重用右列【经“专家-用户”（师生）选订的部分】字与各级“字组”【有汉语思维能力或

选择习惯（即识别、理解、表达或重用“言、辞、链、块”或整体上与之等价于“汉字语素和汉化的词及词组或短语”）即可设订具体的约束条件】。“双列表”左右对应关系用恒等式“$I_D= I_K+I_U$”表示，其中，“I_D、I_K、I_U”分别表示（特定总量的）数据信息与（其中蕴含的）已知信息和未知信息，“$I_D=n^2\approx nm=b$”表示其计算原理（矩阵及线性代数为算法优化和数据结构简化提供了保障）。[3] 图 9 虽只展示了字与词的义项分析(即字间形式信息可标注)途径及可重用工具，但结合图 2 展示的字内与字间形式信息的关系也可理解其具体标注途径。

三、 结语

综上所述，字与词的范畴迭交——汉语组字成语的特点，从（一）字词关系凸显，（二）类的宏观分析，（三）例的微观分析，三方面的探索，可断言“字本位”理应是“言本位”的取向。区分“（以学科和专业知识界定为取向的）专识性义项分析”和“常识性义项分析”，将有利于今后类似的讨论不至于因为相互交流沟通的不畅而造成不必要的误解(会误大事)。也就是说，笔者主张：采用“专识性义项分析”的方式解决概念体系中具有逻辑起点性质的基础概念或术语的歧义问题。如：解决“字本位”的言文误解问题只能在“音字”和“形字”这类“专识性义项分析”之间做出判断或取舍。因为，仅从结构形式而论，作为“汉语基本结构单位的字”的形式化义项只可能有“音字”和“形字”两种结构形式，其中“音字”是基本的（因为它作为言不仅是单音节而且也是语的八个亚类中最小的），“形字”不是基本的（因为它位于形的顶层，它下面少隔着三个低层的偏旁部首结构，最后才是最基本的笔画）。同理，“字本位”的字词以及字素关系的问题也只能在图 4 所示的对比表中来做判断或取舍。因为，这样做是“专识性义项分析”。一旦同时陷入“常识性义项分析”，就必然会出现歧义而可能造成不必要的误解。因为，两种方式所用的标准或尺度经常会是南辕北辙。

为此，本文仅用“专识性义项”解释作为术语之间的相互关系，不涉及“常识性义项”。

参考文献

[1] 许嘉璐.现状和设想——试论中文信息处理与现代汉语研究[J].中国语文.2000 年第 6 期
[2] 马建忠.马氏文通[M].商务印书馆.1983
[3] 刘　勰.文心雕龙[M]. Chinese Text Project[DB]on line.
[4] 黎锦熙.新著国语文法[M].湖南教育出版社.2007
[5] 朱德熙.语法讲义[M].商务印书馆.1982
[6] 朱德熙.语法问答[M].商务印书馆.1985
[7] 徐通锵.语言论[M].东北师范大学出版社.1997
[8] 邹晓辉.义项语汇典例（SVDE）的总量控制模型——人机协作对采用汉语注释的语义词汇典例进行计量分析[A]. Recent Advancement in Chinese Lexical Semantics: Proceedings of 5th Chinese Lexical Semantics Workshop (CLSW-5). Singapore: COLIPS. June, 2004
[9] 邹晓辉.字本位与汉语形式化[M].（徐通锵总主编、潘文国副总主编、杨自俭主编）字本位理论与应用研究. 山东教育出版社.2008 年
[10] 黄昌宁;赵海.中文分词十年回顾[J].中文信息学报. 2007 年第 3 期

注：

* 邹晓辉.字本位与中文信息处理的基础.广东省优秀科技专著基金会推荐与资助出版的专著

** 邹晓辉.一种基于双语自动转换的间接形式化方法（发明专利申请号 2010101752962）