字本位与中文信息处理

——解析“字与字组的关系”探索“汉语形式化”新路

（典型实例：由“一字精解”到“字字精解”）

邹晓辉

0756-5505041 qhkjy@yahoo.com.cn

清华科技园（珠海）融智文化基因工程研究所（筹）

519125 珠海市斗门区井岸桥东恒美花园 15-2栋 201号

【摘要】本文是笔者探索汉语及中文形式化信息处理新方法的总结。英语和基于英语的形式化方法及其好处学界周知，转换生成语法及其后续的各派理论的发展早已为计算机科学和计算语言学普遍接受或了解。模仿它们的汉语词本位、短语本位和句本位理论违背了汉语的特点。因为“汉语中没有词”（赵元任）。“这种跟着西方人思路转的研究是无法实现赶超国际水平的目标的”（徐通锵）。英语形式化方法突破不了中文信息处理的技术瓶颈。如：词的“切分”与“标注”就面临“消歧”难题（俞士汶、孙茂松、黄河燕等）。本项研究课题“摆脱了流行思路的束缚，以字本位理论为基础研究中文信息处理的问题，探索形式化新路。这抓住了汉语特点的关键”（徐通锵），因为“字是中国人心目中的中心主题”（赵元任）。

【关键词】基础语言学，字本位，计算语言学，形式化，计算机辅助，中文信息处理

【专家评语】

“这是一个前沿性的课题。现在语言信息处理的思路大多受国外语言理论的影响，而如何根据汉语的特点，运用信息科学的技术，进行中文信息处理，赶超国际水平，是我们急需探索和解决的一个重大课题。”（本文的）“方向正确，思路清楚，立论有据，是有原创性的新著，其形式化的研究成果也具有广泛的使用价值”。（语言学专家：徐通锵）

“《字本位与中文信息处理的基础——解析“字与字组的关系”探索“汉语形式化”新路》是作者经过长期深入研究和在计算机上通过实践检验的重大科研成果。这个成果的理论意义和实用价值在于：根据汉语的实际特点，运用信息科学先进技术从事中文信息处理，赶超国际水平。”（计算语言学专家：鲁川）

“它较好地实现了与国际学术研究的接轨，因而处于国内同类课题研究的先进水平；作者倡导的融智学新范式和协同智能概念体系，不仅对于我国语言科学和信息科学及其相关学科的研究具有重要的学术探索价值，而且对于建立面向多文种语言信息处理的计算语言数据库和开发拥有自主知识产权的信息产品具有广泛的实际应用价值。”（机器翻译专家：易绵竹）

“语言的形成是一个十分复杂的过程，语言所表达的语义的解析更是一个复杂的问题。本书作者通过对中文语言文字的长期研究，积累了丰富的知识，提出了许多有见地的观点。本文提出了以字为中心，从字出发分析中文语义的一种新的方法。这些思想对于中文信息的自动化处理都提供了一种新的途径。”（计算机科学专家：奚建清）

“（本文）内容新颖，有较高学术水平，…。消解歧义是自然语言处理的关键，本（文）提出的理论和方法，可以对于这个问题的解决提供新的思路。”（自然语言处理专家：冯志伟）

“协同智能计算语言数据库的设计方案中的 13张表很有新意。如果对于汉语的这 13张表一旦建立了起来，那么汉语分析中的各个层次上的歧义就会比较容易地解决。这是一件有创建性的工作。”（清华大学智能技术与系统国家重点实验室专家：苑春法）

引言

长期以来，我们一直缺乏适合汉语及中文自身特点的系统化的语法理论，这严重地制约了中文信息处理的研究进展。《语言论——语义型语言的结构原理和研究方法》（1997 徐通锵）和《基础语言学教程》（2001徐通锵）独树一帜建立了汉语“字本位”理论。本文在此基础之上做了进一步的基础性研究，在尝试对字与字组及其各种关系进行形式化描述的同

时，也尝试对汉语及中文信息处理的形式化方法进行大胆创新。

由本文的标题和副标题可知，“字与字组的关系”的探讨是汉语“字本位”理论关注的基础性问题（属于基础语言学领域）；“汉语形式化”是中文信息处理实践面临的根本性问题（属于计算语言学领域）。两方面结合导致本论题。本文的缘起：北大中文系语言学专家对字的认识分歧（至今尚未达成普遍一致的共识）。试问：作为自然人的专家尚且无法消除的分歧，怎么让计算机系统去重用？这类性质的难题也是主张强人工智能观点的中文信息处理专家们所面临的棘手问题。如，中科院计算机语言工程研究中心机译专家就说：对机器翻译而言，只有一个难题，就是消歧。清华大学计算机系自然语言处理课题组专家也明确地指出汉语在“分词”与“标注”上存在技术瓶颈。北大计算语言学研究所专家还十分明确地指出（汉语及中文的）形式化非常困难。中国社科院语言学研究所机译专家公开指出语言学理论滞后制约了中文信息处理技术的发展。

同样研究自然语言，不同的学科有不同的视角，普通语言学站在人类智能主体的立场，采用的是自然人的视角；计算语言学站在人工智能代理的立场，采用的是计算机的视角；工程融智学站在协同智能计算系统的立场，采用的是自然人和计算机两者交互协同的视角。本文就是对从（必然兼容且优于前两种视角的）第三种视角而提出来的研究课题的回顾。

工程融智学的方法，以人机“合理分工、优势互补，高度协作、优化互动”的方式独辟蹊径，提出了自然语言理解的工程模型（基于 Z-ASCII 的 GTCM/STCM与基于 Z-Unicode 的 GSCM/SSCM）及应用模式（SDVE），如：“两典一册”。部分成果（1997-2005）已得到学术界多个课题组专家们不同程度的认可（见：专家评语）。本文将重点介绍其中近期取得的进展。汉语“字本位”理论方面，本文明确表述了字的迭交原理，直观地表述了字与词两种思维模式，二字组的构造原理；中文信息处理方面，本文明确给出了字处理的“三合一”设计方案（经过“中文计算机输出输入系统”、“终极标准信息交换码”和“大字符集可小字符集化的字型库”协同试运行一段时间之后可以中文基因芯片的形式固化），提供了“两典一册”（经过“合作型生产式教学法”推广普及活动检验之后可以中文语法芯片的形式固化）的示例。全局形式化标准平台，可为“中文基因”和“中文语法”信息的提取以及“（汉英/英汉）双语概念及命题”知识的提取，提供人机“合理分工、优势互补，高度协作、优化互动”的优化环境。从而，可进一步为“中文基因芯片”和“中文语法芯片”以及“（汉英/英汉）双语知识（概念及命题）芯片”的封装奠定形式化基础。这涉及业内普遍感兴趣的一组关键问题的解决，是适合汉语字本位语法形式化表述进而可改观中文信息处理形式化现状的新方法。

正文

“字与字组”的完全形式化解析，主要有三个步骤，即：1、字内信息处理（音形层解）；2、字间信息处理（音节串解）；3、字外信息处理（义项分解）。具体操作，涉及：人机之间的“合理分工、优势互补，高度协作、优化互动”。目标是：由“一字精解”到“字字精解”。

1 字内信息处理 [音字与形字（含：笔画与偏旁部首）的“层解”]

图 1是“线串型结构”与“层面型结构”虚拟“层解”示意图。

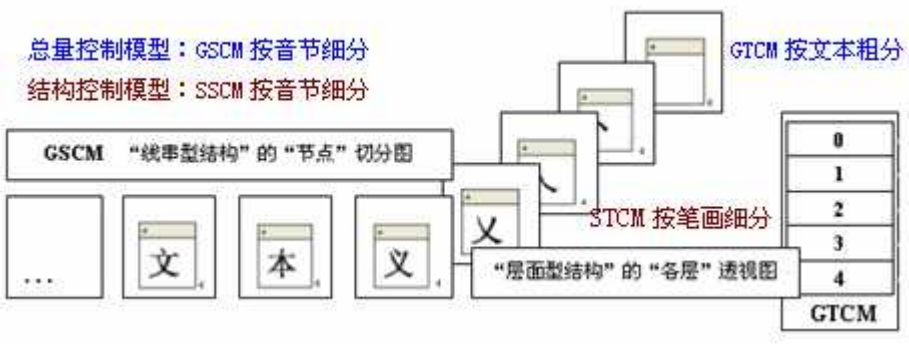


图 1

笔者认为：汉语的单音节的字与英语的混音节的词是两种语言形式系统最根本的区别。为此，本文提出音字的概念和混音节的称谓，把“形字（层面型结构）、音字(线串型结构)、实字、虚字、用字、解字”并列，旨在突出：音字与字音，混音节与多音节，形字与字形、用字与解字、前字与后字、(二字）释辞与（一般）二字组的区别，强调：各自关注的重心、焦点以及主要研究对象的不同。本文之所以采用部分新概念和新称谓是为了使表达更到位，同时，也是给字的迭交原理、释辞公式和语块方阵的介绍等做必要的准备。

由图 1直观地展示虚拟优化字库中线串型结构的音字与层面型结构的形字被层解模型。它体现字在形式上是由音字与形字这两个类“迭交”的复合类。可称之为字的形式迭交。

图 2是字内信息解析示意图。

图 2

（虚拟的）音字和形字“迭交”不仅反映字的音形关系，而且还反映一字之内有丰富的字形信息，如：笔画和偏旁部首。粗分有五个层次，即：笔画、缺省的字内字、变形的字内字、正形的字内字、字；细分则有多个层次，即：一、二、三、…多个笔画。这是字内信息自动计算的基础。

音字和形字“迭交”所揭示的音形关系，既可由“音字串”的“串解”与“形字层”的“层解”直观展示（见：图 1-2音字和形字“迭交”示意），也可由“音字顺序编号”与“形字结构代码”的双列表精确记录（见：图 3）。

图 3是字内信息“层解”示意图。

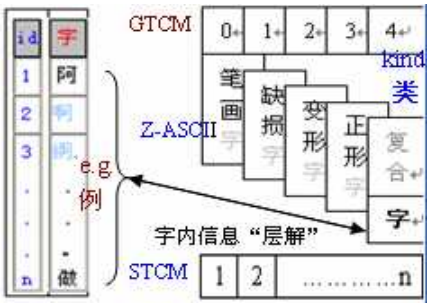


图 3

通过图 3 可揭示每个字在“文本总量控制模型”（GTCM）和“文本结构控制模型”（STCM）中的特定序位。在图 3双列表中的一个个字例是位于 GTCM 第 4进阶的新型字典里可“层解”的字。GTCM 第 1-3 粗分进阶记录字内偏旁部首编号信息；STCM 第 1-n 细分进阶记录字内笔画编号信息。

计算机系统通用的显示字库（Font）是依据计算机内码（如：GB2312 ,GBK和 GB13000.1 汉字信息交换码）的国家标准及国际标准（即：Unicode国际统一的字符编码标准）排序的。由于每种字库均由点阵曲线加工的模拟数据记载（占据很大的存储空间），因此，受到存储空间和开发成本双重限制。本文介绍的字内信息处理的“层解”方法的应用，可大幅度缩小字库的存储空间，并可显著降低字库开发成本，而且还容易改进甚至再造或重构各种字体。

2字间信息处理（音节串解）

图 4是“线串型结构”的“串解”示意图。

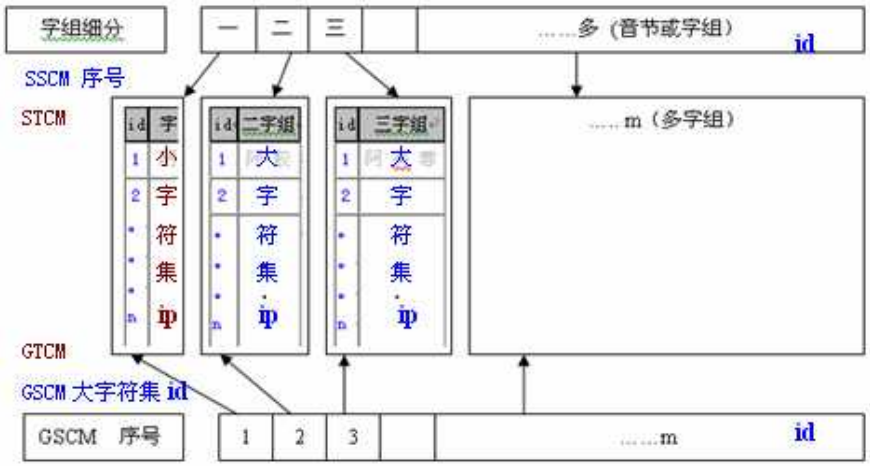


图 3

通过图 4可揭示每个音字在“音节总量控制模型”（GSCM）和“音节结构控制模型”（SSCM）中的特定序位。由于中文形字与汉语音字以及（音字）字组分属“层面型结构”和“线串型结构”两个系列，因此，基于笔画表编号 id的小字符集实际位图 ip与基于音字表编号 id的大字符集虚拟位图 ip可通过字表与字组表的前后台构成具体的函数关系。

文本控制模型与音节控制模型在图 4所示的双列表中是合二为一的。其中，音字一览表就是图 3所示的位于 GTCM第 4进阶的新型字典里可“层解”的“音、形“迭交”的字例”双列表，二、……、多（音字）字组一览表均为其两两、三三、…、多多轮排的采集记录。图 4蕴含：认知科学、逻辑学、数学、计算机科学等相关领域的科学原理和技术方法。

认知科学原理：理解，实质上是一种识别关系的能力。其特点有二，a、对关系的识别；b、对“问题状况”形成一种“内部表示”。各种“问题状况”涉及“语义丰富领域”（对关系的识别）/模式识别。各种“问题状况”的“内部表示”涉及局部理解/知识的获取与表达。对（语言）关系建立的“内部表示”涉及全局理解/系统的知识表达。

例如：文本总量控制模型和音节总量控制模型就是 GeneCulture这个具体的智能主体“对‘（语言）关系’识别”以后建立的“内部表示”（静态模型）。其中，包含：各种“问题状况”的“内部表示”（动态模型）。这个模型及其实施例，是否被其他不了解它的具体智能主体“发现”或“认同”，则有待进一步的实践或共享/重用之后做出新的评价或评估，不过，现在我们的实验和分析证明它具有可计算、可操作、可重用、可共享的特征（故得到了部分具体的智能主体“发现”或“认同”，因此，进一步的实验和推广活动可进行下去）。字与（各级）字组（形式）的关系，在上述认知模型中以“潜在”（理想状态）和“显在”（受限状态）两种方式被记录在案。通过“三化”加“三注”等具体“限制方式”，我们可针对“目标用户群”从中选取相应的“义项字典”与“字组用例”，作为构建：数字化、标准化、高性能的各种标准化与个性化统一的实用语汇工具（含工具书）为汉语教学和中文信息处理提供计算机辅助（CA）。

逻辑学原理：在图 4所述“字与字组划分数字化”模型中，不仅一、三、…、多序号与 1，2，3，…，m序号之间同义并列，而且，一、三、…、多字组（音字与音字串）序列与 1，2，3，…，m 数字（代码）序列之间也同义并列。根据“同义并列，对应转换”公理/“序位逻辑”法则（通则），任何两个形式信息体系，一旦“同义并列”，即可“对应转换”。另提示：“义项数量与字组长度之间的反变关系”与“内涵与外延之间的反变关系”同理。

数学原理：在图 4所述数字化模型中，由各表序号（m）和各表中同义并列的数字与文字的行的序号（n）构成的矩阵序列，即：线性方程组（$\Sigma a_{mn}x_n = b_m$）常数项序列。

计算机科学原理：在图 4所述“字与字组划分数字化”模型中，由各表序号（m）和各表中同义并列的行的序号（n）构成的矩阵序列，等价于计算机数据（仓）库的（一系列）表的序号（m）和各表中行的序号（n）构成的矩阵序列。在计算机关系数据库各表的前台直接呈现以及后台间接计算的字与（各级）字组（形式）与后台直接计算的数值（数字或代码）之间，不仅是同义并列的逻辑关系，而且，也是一一对应的函数关系。

在计算机标准化形式体系与（各个）自然人（实际选择或使用的）个性化内容体系之间可构成这样一种协同智能计算关系，即：人工与自然语言处理兼容的可计算、可操作、可重用、可共享的认知模型。

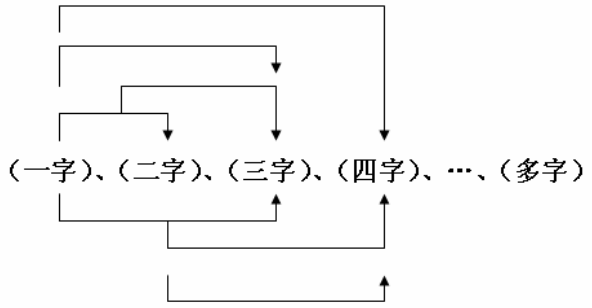
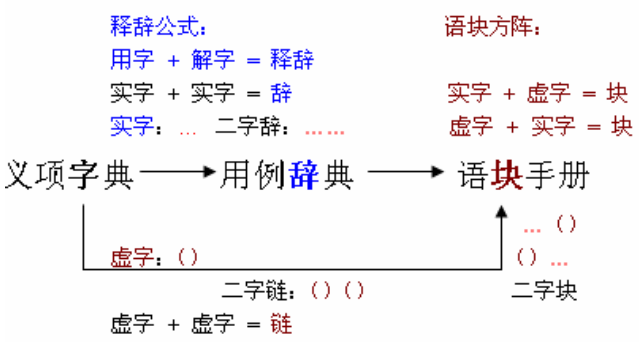


图 5是字与字组关系示意图。图 5通过图 5 可揭示音、形“迭交”的字与二、……、多（音字）字组的关系及其具体的组分机理。这是字为什么可充当汉语及中文的基本结构单位的原

因。基于此，电脑芯片只要存储图 3所示 Z-ASCII 笔画编号 id和字的编号 id与图 4所示小字符集编号 ip与大字符集编号 ip就可再现字和各级字组，以便计算机系统和自然人用户有针对性地重用。

图 6是语法原理示意图。 图 6



通过图 6 可直观揭示字与各级字组尤其是二字组关系中所蕴涵的中文语法基本原理，即：释辞公式，语块方阵，语法探针或链。三合一可构成中文语法框架的基础，同时，也就是中文字间信息处理的原理。因为，多字组可由字与二字组衍生。中文标点符号信息处理也可在上述中文语法框架的基础（字与二字组的关系）上发展起来。二字关系及原理可指导“两典一册”的提炼或编撰；“两典一册”的完善，也可促进中文语法原理体系和中文字间信息处理方法体系的完善。

3 字外信息处理（义项分解）

图 7是两点论与三段论（两种基本的思维方式对比）示意图。

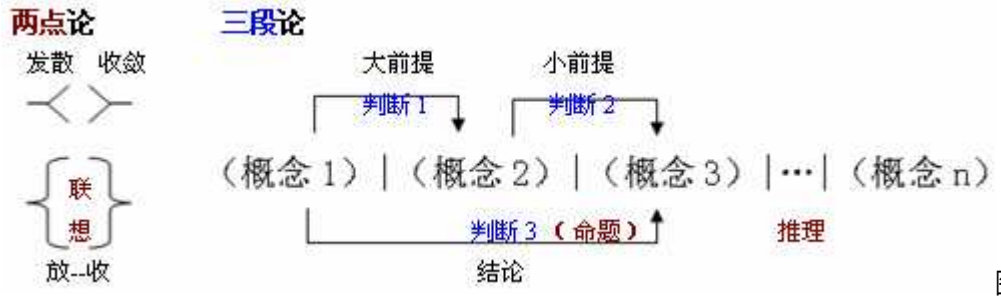
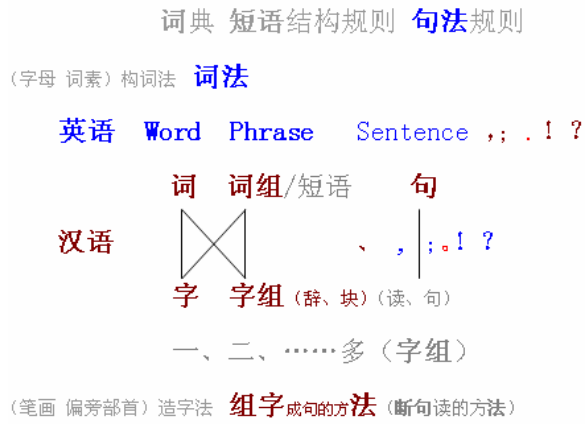


图 7

通过图 7 可揭示汉语及中文基于字的“两点论”与英语及英文基于词的“三段论”两种基本思维方式的区别。“两点论”(图 7) 与“二字关系”(图 6) 可体现汉语及中文思维方式及辞语形式有机结合的基本特点。多义项的字由发散（生歧）到收敛（消歧）可通过二字组（无论是辞还是块或是链）这一基础性环节而在相应的概念群、辞语族及其关系链中实现。明确表示概念的辞，因为，免除了独字词的多义性，所以，更适合基于“三段论”的推理。

图 8是字词关系示意图。 图 8

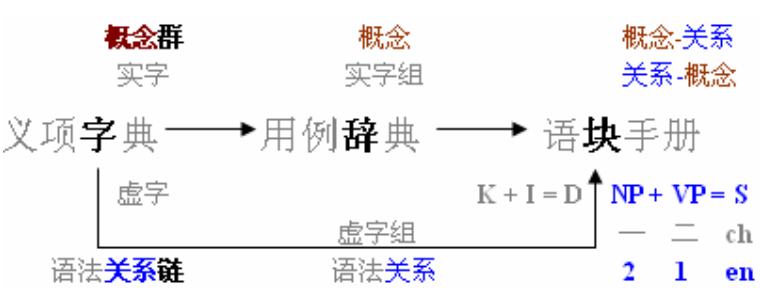


通过图 8可揭示汉语及中文的字和英语及英文的词所体现的两种语义体系及语法架构的区别和联系。“单独存储模型”可记录字与词的区别；“共同存储模型”可记录字与词的联系。图 8展示了笔者的（汉英 /英汉）双语观及相应的协同存储模型，即：字在英语中可表现为词与词组，词在汉语中也可表现为字与字组。

此观点及模型可指导双语实践。在定性及定量分析的基础上可提炼出汉语及中文的形式化科学原理和“两典一册”(即：《义项字典》、《用例辞典》和《语块手册》)。

同时，也可相应地提炼出英语及英文的《义项词典》、《用例词组》和《短语手册》。进而，还可提炼出常用 (汉英 /英汉) 双语的数字化教学用 /日常查询用 /信息处理用的工具书。

图 9 是链群示意图。图 9 与实字联系的概念群和与虚字联系的关系链，可以把一个实字与一群概念，一族辞与一群概念，相联系；一个虚字或虚字组与关系链，一族块与一系列关系及一系列概念，相联系，可通过图 9揭示。

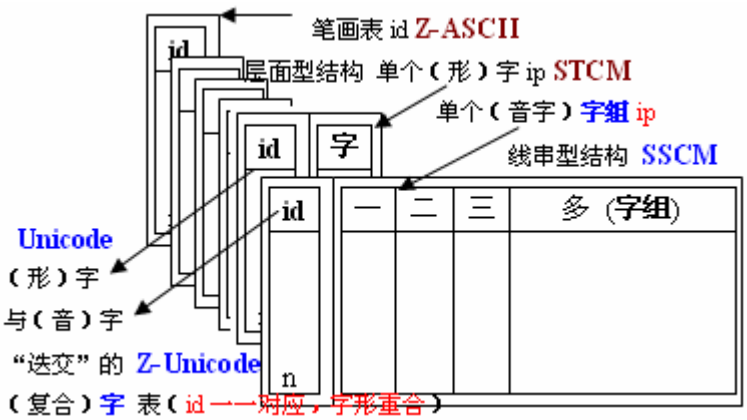


实字蕴含或牵动的概念群（多义项“迭交”）可由含该实字的多个辞（实字组）表达的多个概念体现（多个辞“迭交”于该实字）。虚字蕴含或牵动的关系链可由含该虚字的多个块[（实虚/虚实）字组]表达的多个关系-概念/概念-关系体现（多个块“迭交”于该虚字）。

汉语及中文关注充当话题的辞语（如：英语及英文的主语）；英语及英文关注充当谓语的动词（如：汉语及中文的说明）。由图 9 所示的短语结构与句子等价关系式 NP + VP = S 及其下方的一、二和 1、2 标号可见两者关注重心或焦点有不同语序。如果把 NP + VP = S 视为信息学基础研究的语义信息公式 K + I = D 的特例，则英语及英文关注谓语（动词性短语）动词的未知语义信息胜于关注已知的（名词性短语）主语；汉语及中文关注话题（名词性短语）辞语的未知语义信息胜于关注已知的（动词性短语）说明。

4 标准平台　　图 10

图 10是原理示意图。



由图 10可揭示字库的层面型结构编号（ip）是笔画编号（id）层解记录，字组库的线串型结构编号（ip）是字的编号（id）串解记录，在此标准平台上分别由 Z-ASCII 编号（id）与 Z-Unicode 编号（id）可确定（复合）字结构编号（ip）与字组结构编号（ip）的记录。其前后台数据结构特点和人机协同操作原理及方法，可由表格化、数字化、字组化三个基本步骤（简称：三化）具体表达。

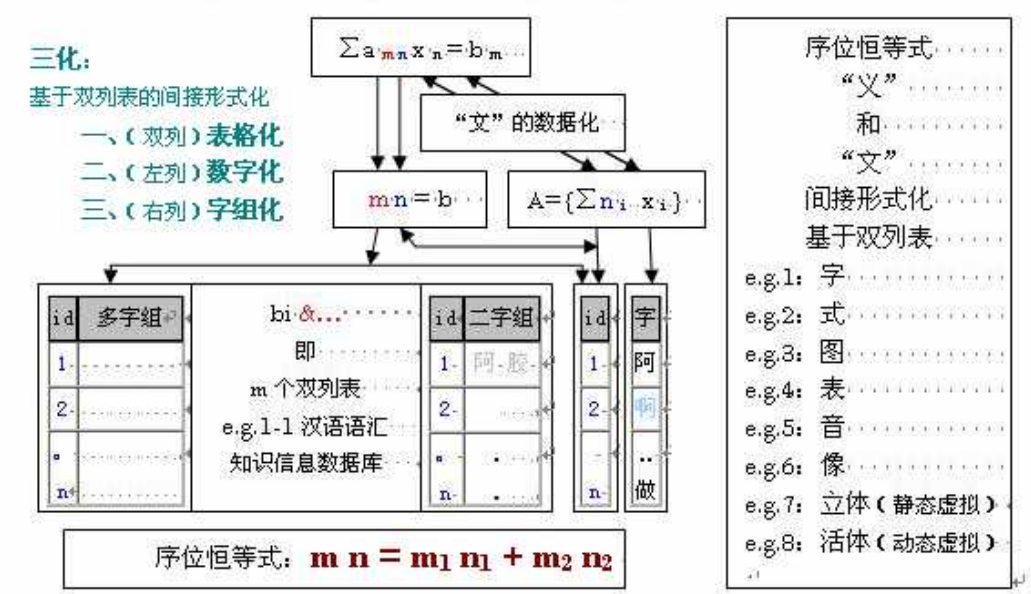
标准平台的表格，有三种结构类型，即：单列表、双列表、多列表。简称：三表。

“三表”的前后台数据结构，兼顾自然人（人类智能）和计算机（人工智能）两方面的特点，强调人机之间“合理分工、优势互补，高度协作、优化互动”的智慧融通和融通智慧（智融和融智），是协同智能计算系统定制的数据结构。单列表是只有一种数据类型的电子表格。其特征在于：该列数据“异义排列，序趣简美”。它可以是任何一种数据类型，但只能是其中的一种。对双列表而言，它要么是左列，要么是右列。对多列表而言，它只能是其中的一列。双列表是具有两种数据类型的电子表格。其特征在于：两列数据“同义并列，对应转换”。也就是说，标准平台的双列表遵循“信息基本定律”安排左右两列数据的前后台数据类型。多列表是可有多种数据类型的电子表格。其特征在于：多列数据“经纬阵列，唯一确定；多维选列，非非各平（即：非同步、非对称、各自平衡）”。“三表”分用（编号表对电脑/字组表对用户）合用（编号表与字组表以及相应的知识表均对开发人员）均可。

标准平台的数字，有三种集合类型，即：单一集合、分层集合、标志集合。简称：三集。笔者发现：如果说“单一集合”是“还原论”的科学基础，“杂多集合”是“整体论”的现实基础，那么，“标志集合”和“分层集合”则是“（各门学科及其知识可做到）各就各位论”的科

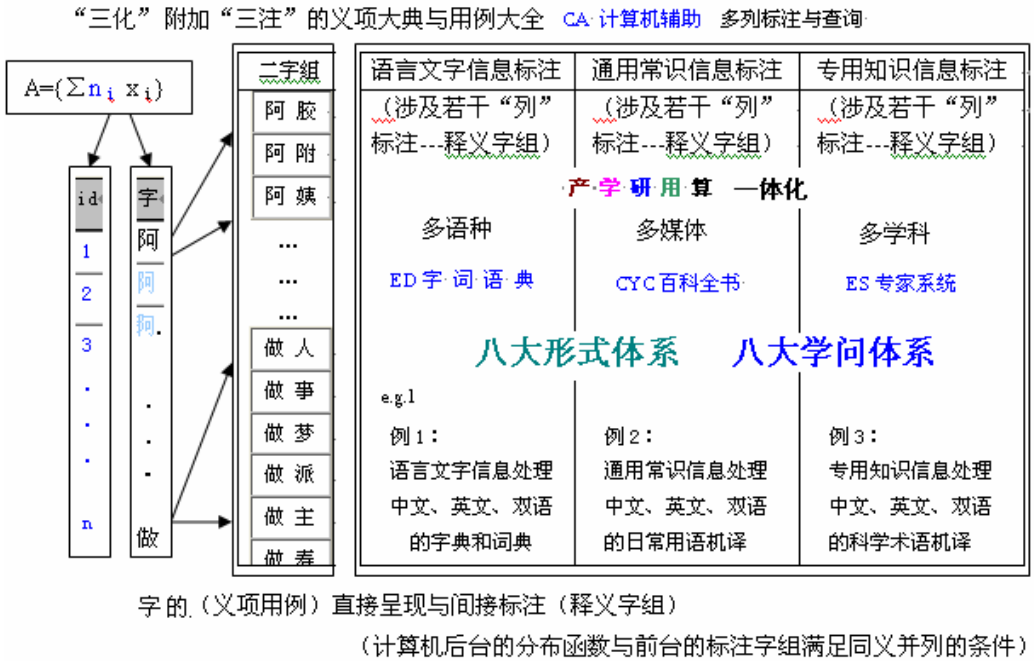
学基础。笔者划分的单一集合、分层集合、标志集合，在数学上分别对应于“集合与映射”原理中所说的集合、集合的直积、商集。至于，日常生活中的集合，可称之为：杂多集合。这样划分的好处，可举例说明：仅就形式集合而论，笔画表属于单一集合，形字表属于分层集合，音字表属于标志集合，复合字表属于杂多集合。处理字与字组的标准平台，也是提炼“两典一册”的形式化通用平台。

图 11是提炼“两典一册”的形式化通用平台“三化”示意图。 图 11



通过图 11可揭示提炼“两典一册”的通用平台的科学原理和操作界面。《义项字典》与《用例词典》以及《语块手册》的素材，均可取自图 14所示的 m个双列表分层集合。在此，形式化是通过“三化”的步骤而落实的。其中，字与二字组的关系可作为语言文字形式信息提取与语言文字内容信息提取或知识表达提供有效的格式化操作示例。

图 12是提炼“两典一册” 的形式化通用平台“三注”示意图。 图 12



字的（义项用例）直接呈现与间接标注（释义字组）

（计算机后台的分布函数与前台的标注字组满足同义并列的条件）

通过图 12 可揭示字组标注的三种类型，即：语言文字信息标注、通用常识信息标注、

专用知识信息标注。简称：三注。如：可通过笔画表、形字表、音字表、复合字表而实现对字的形式信息标注；可通过实字表和虚字表而实现对字的内容信息标注。

5 两典一册

由哲学到数学再到计算机科学的探索进程中逻辑学的发展涉及了概念化和形式化两个重要方向。因自然人与计算机各有所长，故融智学采取了人机互助互补的观点、策略和方法。“两典一册”的设计直接应用了融智学的主张和工程融智学的具体做法。

首先，利用自然语言处理的文本与音节两组模型，把汉语及中文的基础结构形式化。图 13是基础结构与两组模型示意图。

人/机两用	Z-ASCII 笔画（表）
G/STCM Z-Unicode 复合字（表）	（文字：部首）形字
G/SSCM	（语音：音节）音字
《义项字典》	（语意：概念）实字
《用例辞典》	（语法：关系）虚字
《语块手册》 字组（表）	（语用：搭配）字组

图 13

Z-ASCII 笔画（表）与 Z-Unicode复合字（表）是汉语及中文形式化的基础。因为，（形）字的计算，基于笔画表（Z-ASCII包容且兼容 ASCII），（音）字及（音）字组的计算，基于（音形“迭交”的复合）字表（Z-Unicode包容且兼容 Unicode）。字的形式化是其中承上（“层面型结构”）启下（“线串型结构”）的关键步骤。G/STCM和 G/SSCM是适合人/机的模型。

《义项字典》解释（形/音）字的义项，分“实字”和“虚字”两种类型。其中（形/音）字的义项“层解”通过“STCM和 SSCM两表”区分形字和音字而实现。（形）字的义项“层解”通过（形）字表，标注笔画和部首而实现。（音）字的义项“层解”通过（音）字表，标注多音字及其用法而实现。（音）字的义项“串解”通过字组表，切分字组而实现。实字的义项“分解”通过辞表，切分实字而实现。虚字的义项“分解”通过链表，切分虚字而实现。（形/音/意义）综合“迭交”的字由各表“分解”。

图 14是字的综合“迭交”示意图。

图 14

通过图 14 可揭示语言学主要分支学科（如：语音学、文字学、语义学、语法学、语用学、字典学）的一组微观研究对象（如：音字、形字，实字、虚字，用字、解字）。汉语及中文“字内、字间、字外”信息处理，以字为焦点，涉及字的综合“迭交”。其中，形字“层解”可获得：字内（小字符集笔画顺序编号和大字符集偏旁部首顺序编号）信息，此处不计“形字与音字”的“迭交”信息；音字“串解”可获得：字间 [大字符集（音字与形字迭交复合的）字的在字表中的顺序编号和在字组中的结构编号]信息，此处不计“普通话、方言、古字音”的“迭交”信息；实字义项/概念群“分解”可获得：字外（辞与概念的顺序编号）信息，此处暂不计“实字与虚字”以及“虚字与关系链”的“迭交”信息。仅就形式而论，音形迭交的音字（如：形声字的声符也就是 1325个音字）比音形分离的拼音与形字更具汉语的特性。用字或解字涉及对字意的应用或解释。

接着，利用字与二字组的关系，融“形式化”与“概念化”为一体，可通过形式化保证计算机自动识别和精确重用，同时，可通过概念化满足自然人正确识别和准确理解乃至恰当表达（有针对性地重用）。如：视字为对象语言，视二字组为元语言，在网络与计算机辅助环境或条件下，由“一字精解”到“字字精解”。

图 15

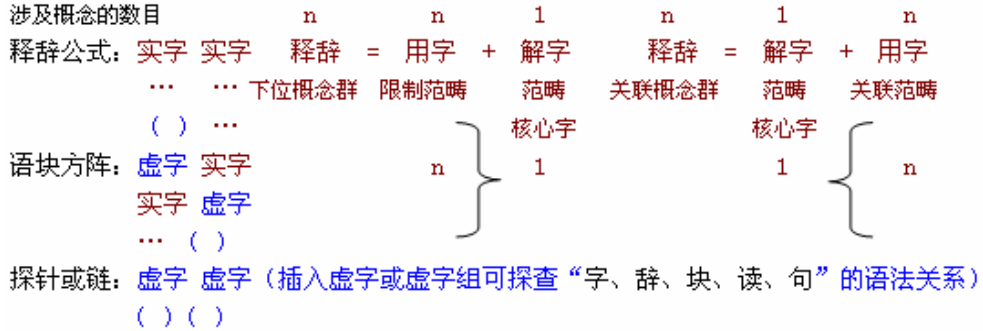


图 15二字关系

在“两典一册”中具有枢纽地位的二字关系。释辞公式反映实字与实字的线性组合；语块方阵反映实字与虚字，虚

字与实字的阵列组合；探针或链反映虚字及虚字组具有揭示语法关系的功用。虚字或虚字组（作为语法关系链，省略时人为地插入可像探针那样具有检测并凸显其语法特点的功用）与实字或实字组/辞结合，组成（语）块（加逗号则为读，加句号等则为句）。探针或链，一旦归纳整理成册就可系统地展示汉语及中文语法的全貌。因此，可以说，实字是语意学的字类，虚字是语法学的字类，用字是语用学的字类，解字是字典学的字类。作为核心字的解字的（类/范畴/基本概念）定义或解释，可统帅二字组——包括：二实字构成的释辞族；二虚字构成的链；一实字和一虚字或者一虚字和一实字构成的块。实字、虚字，用字、解字，属内容方面，音字拼音标注和形字笔画及部首标识属形式方面。

进而，逐步完成字与（各级）字组（关系）的详解（以备网络与计算机辅助教学、科研、生产、日常应用、自动处理或由针对性地重用）。

《用例辞典》解释（实字）字组/辞，分“用字”和“解字”两种类型。其中，释辞族，由两种情况，即：由前字充当解字的相关概念群（如：字音、字形、字意）与由后字充当解字的下位概念群（如：音字、形字、实字、虚字，用字、解字）。在此，解字就是核心字。

《语块手册》解释（虚字）字组/链，分“近距”和“远程”两种类型。其中，虚字与虚字直接组合，为“近距”接续（如：当且仅当）；虚字与虚字间接组合，为“远距”接续（如：不但…而且…）。

图 16

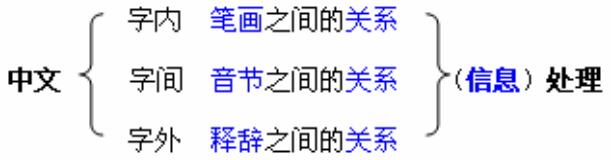


图 16是中文信息概要示意图。

汉语及中文“字内、字间、字外”信息，表示：笔画之间、音节之间、释辞之间的关系。如果说自然人识别、理解、表达的则是部首之间、音节/字之间、辞之间（概念化的）关系；那么，就可以说计算机识别和处理的实际上是笔画之间、音节/字之间、字组之间（形式化的）关系。因此，可用（作为元语言的）二字组（释辞）解释（作为对象语言的）字。也就是说，“两典一册”所记录的就是这一系列的关系/信息。

普通的字典、辞典、语法手册，是供自然人查询用的；经过计算语言的电子字典、辞典、语法规则库，主要是供计算机（如：中文信息处理和机器翻译）查询用的。计算语言学借助程序语言来处理自然语言。国外的形式语法及语义/意和自然语言的形式化方法，对汉语及中文信息处理（如：各种各样的加工——“标注”），虽然取得了一系列成果，但是却遭遇了“分词”的困难——这是汉语及中文信息处理的一个根本性的技术瓶颈。

实践证明“这种跟着西方人思路转的研究是无法实现赶超国际水平的目标的”（徐通锵）。英语形式化方法（何况英语自身的形式化问题还没有解决）突破不了中文信息处理的瓶颈。如：中文词的“切分”与“标注”就面临“消歧”难题（俞士汶、孙茂松、黄河燕等）。

必须指出：对人机两用的《义项字典》、《用例辞典》、《语块手册》不存在“分词”问题。因为“词”已被（“两典一册”清楚表示的）“字、辞、链、块”所取代了。如：（形）字的计算，基于笔画表（Z-ASCII），（音）字及（音）字组的计算，基于（音形“迭交”的复合）字表（Z-Unicode）。辞的计算，基于实字表；链的计算，基于虚字表。块的计算，基于“字表、辞表、链表”（D = K + I包容且兼容 S = NP + VP）。

最为关键的是“汉语中没有词”（赵元任），“字、辞、链、块”是人机两方面都可识别和处理的。汉语及中文的形式化的字本位理论和概念化的自动查询工具（标准平台）的功用，不仅符合汉语及中文的思维及表达习惯，而且，也具有与英语及英文的思维及表达习惯兼容的通道。汉语及中文的“字、辞、链、块”可取代“词”（西方人心目中的中心主题）而“字是中国人心目中的中心主题（赵元任）”，因此，“摆脱了流行思路的束缚，以字本位理论为基础研究中文信息处理的问题，探索形式化新路。这抓住了汉语特点的关键”（徐通锵）。

6自然语言处理与理解的基础

基于 Z-ASCII的 G/STCM和基于 Z-Unicode的 G/SSCM 是自然语言处理与理解的基础。

其中，STCM和 SSCM是适应计算机处理的后台模型；GTCM和 GSCM（简称：两表）是适应自然人理解的前台模型。前后台之间是等价的。

“两表”及其作用（“三化”和“三注”）概要简述。

文本总量控制模型（GTCM）						
分表	标点	进阶层式	汉语	拼音	英语	标点
1		0	笔画字 基本笔画	字母	字母	
2		1	损形字 偏旁部首		词头和词尾	
3		2	变形字 偏旁部首		前缀和后缀	
4		3	字中字 偏旁部首		词根	
5	顿号	4	“字”（形字音字“迭交”融合）	单音节	（混音节）词	逗号
6	顿号	5	“辞”（全由实字构成的多字组）	多音节	（多音节）词组	逗号
7	顿号	6	“块”（附加虚字构成的多字组）	多音节	（多音节）短语	逗号
8	逗号	7	“读”（表示：语气上的停顿）			逗号
9	句号	8	“句”（表示：语义上的停顿）			句号
10	（提行）	9	“段”（具有：段意） （分层）			
11	（题名）	10	“篇”（具有：主题） （分节）			
12	（分篇）	11	“册”（涉及：文集和书库）（分章）			
13	（分册）	12	“集”（涉及：书库和数字化图书馆）			
			字本位（形字、音字、实字、虚字）			

图 18

图 18是 GTCM（粗分模型）示意图。

通过图 18可揭示汉语及中文一览总表的 13个分表所表达的 13个类。其中，各分表均采用“双列表”，即：左列，前台是十进制数，后台是二进制数；右列，前台是自然语言（汉语及中文）符号，后台是二进制数。数据类型分别是整型/自动编号（id）、逻辑型、字符串、超级连接/序位编号（ip）。左、右列，前、后台，均“同义并列”。由此建立“三化”汉语及中文形式体系。本文重点探讨“字、辞、块”3个分表，其它 10个分表仅作简要介绍。

对汉语及中文而言，GTCM的 0，1，2，3，4五个分表，所记录的“字内信息”易“计算、操作、重用、共享”且“符号”总量有限。

GTCM的 4-6三个分表与 GSCM的 1-m个分表，所记录的“字间信息”在总量上完全相等，不同的是前者 GTCM需人助机识别而后者 GSCM计算机可自动识别，即：易“计算、操作、重用、共享”且各个分表的“符号”总量也有限。

GTCM的 7-12六个分表，所记录的“字外信息”的粗分和细分均受到具体目标用户日常处理能力的限制，其总量也限制（由于这部分不是本文的探讨范围，故省略）。

这个自然段的术语，仅供有兴趣学习或参与“协同智能计算系统”设计的读者参考——其他读者可不读此自然段。GTCM的 0分表被命名为子全域/单一集合，其中的元素数目十分有限，是构成 GTCM的 1-12个分表的所有被命名为超子域的元组的基因文本或文本基因——因其特性类似生物基因(ATGC)而得名。GTCM的 0-12个分表被命名为 13个进阶。基因文本元素是子全域与超子域的连接纽带。超子域的复杂程度，在微观上取决于元素的分布状态（一旦跨进阶或跨模型就会出现“迭交”情况——如：GTCM的第 4个分表和 GSCM的第 1个分表），可解析，具体的计算、统计、分析，视具体的子全域而定；在宏观上视具体的进阶（如：细分形式体系的各个一览表）而定。超子域可解析程度视组配结构的复杂程度或迭交情况而定。

单音节与混音节的区别和联系

同样位于 GTCM第 4分表（图 18）的（汉语）（音）字与（英语）（单）词，却有显著

的区别。一方面，就文字结构（形式）而言，与（音）字发生“迭交”的（汉文）（形）字，是：基于笔画的“层面型结构”；（英文）词（形），是：基于字母的“线串型结构”。另一方面，就语音结构（形式）而言，与（形）字发生“迭交”的（汉语）（音）字，是：单音节；（英语）（单）词，是：混音节（含：单音节、双音节、多音节）。

字、辞、块与词、词组、短语

同样位于 GTCM的 4、5、6 分表（图 18）的（汉语）字、辞、块与（英语）词、词组、短语之间的微妙关系，仅从“辞、块”与“词组、短语”这种“粗分”形式是难以发现的。

探讨字、辞、块，即：GTCM的 4，5，6 三个分表，是图 18 的焦点。这里（图 19 将对“辞、块”做进一步细分，以便进行计算机自动化处理）对汉语与英语的比较，强调：单音节的字与混音节的词的区别。

音节总量控制模型（GSCM）

分表	有、无“标点”	进阶层式	汉语	拼音	英语（词）	英语（词组或短语）
1		1	字	一音节	（一音节）词	
2		2	二字组（辞或块）	二音节	（二音节）词	（二音节）词组或短语
3		3	三字组（辞或块）	三音节	（三音节）词	（三音节）词组或短语
4		4	四字组（辞或块）	四音节	（四音节）词	（四音节）词组或短语
5		5	五字组（辞或块）	五音节	（五音节）词	（五音节）词组或短语
M		M	多字组	多音节	（多音节）词	（多音节）词组或短语

图 19

图 19是 GSCM（细分模型）示意图。

通过图 13可揭示汉语及中文一览总表的 1-m 个分表所表达的 1-m 个类。其中，各分表均采用“双列表”。这里的汉英比较仅限于音节形式。汉语语汇的细分就是以字为“线串型结构”的起点或节点作为度量（各级）字组的基本尺度（基本结构形式单位）。两表（图 18-19）结合使用，可突出形字与音字的“迭交”关系。GSCM突出音字。请注意：这是仅仅就汉语而论，不涉及汉语拼音。如果说混音节与多音节在 GTCM中仅显现出称谓不同，那么，在 GSCM中就表现出了实质差异。其区别在于：是否包含单音节？图 14可使答案一目了然。

（汉）字与（英）词比较

由图 19 可见（汉语）（音）字与（英语）（单）词在 GSCM中的地位是不同的，这两种语言各自的基本结构（形式）单位的区别是显而易见的（形态上也有不同）。第一，静态区分：字仅仅限于 GSCM的第 1 个分表，词则可位于 GSCM的 1-m 多个分表。第二，动态区分：字占据的节点（含：起点）是单音节，词占据的节段（含：节点）是混音节，两者都属于“线串型结构”。

GSCM揭开“微妙关系的神秘面纱”

前面提到的（汉语）字、辞、块与（英语）词、词组、短语的微妙关系，在 GSCM和图 19 中得以展现——被揭开了“微妙关系”的神秘面纱。具体要点如下：

首先，就语言结构（形式）而论，字与词之间的区别是最根本的。因为，同样位于 GSCM的 2-m 多个分表的（汉语）字组（辞或块）与（英语）词组或短语之间的区别，皆由作为其基本结构单位的字与词的区别而派生——各自的组配法则也有相应的区别。

其次，（英）词的混音节（含单音节、双音节、多音节）与（英）词组或短语的多音节（不含单音节）的（形式）区别，也显而易见。除此之外，两者的区别还有：前者是构成后

者的基本结构单位，而反之则不能成立；构成词的音节之间无空格，而构成词组或短语的词之间有空格——计算语言学界认为这是英语的一个优点，增加了英文信息处理的识别标识。

最后，如果仅仅限于 GSCM来看，那么，很显然，构成（汉语）字与字组的音节（音字）之间无空格。计算语言学界认为这是汉语的一个弱点，增加了中文信息处理的困难。但如果考虑 GTCM可提供的字内信息，特别是考虑：字内和字间信息的综合利用，那么，中文信息处理应当有自己的优点（这在后面将会有进一步的分析）。

区分字与词的工具

正如 GTCM和 GSCM可帮助我们区分形字与音字一样，GSCM和图 14也可帮助我们区分：（音）字与（英）词以及在音节关系上认识（英）词与（音）字、辞、块的关系和（英）词组或短语与（汉）辞、块的关系。

GSCM奠定理论分析和实践处理的基础

GSCM展示（作为汉语结构的）（音）字与（各级）字组的（形式）特点，为进一步提炼（汉语）（音）字的形式化定义以及（各级）字组数字化分类（或划分）奠定了基础，即：表格化、数字化——意味着：可计算（可统计、可分析）、易操作。

字与字组的关系——兼谈字与词的区分

字与（各级）字组的关系（基础是：字与二字组的关系），涉及：形式与内容两方面。

三化

从形式方面看，字与（各级）字组的（形式）关系，可借助“两表”实现：字的定义形式化，字组划分数字化，义项呈现字组化（即：三化）。这是实现：计算机辅助（CA）处理汉语形式的一条捷径。

三注

从内容方面看，字与（各级）字组的（内容）关系，可借助“两表”进行：语言文字信息标注，通用常识信息标注，专用知识信息标注（即：三注）。这是实现：计算机辅助（CA）处理汉语内容的一条捷径。

奠定“中文信息处理”的系统工程基础

基于“两表”的“三化”加“三注”，从形式与内容两方面对“字与（各级）字组的关系”给出了静态的系统描述（相当于：在“现实需要”与“理想目标”之间架设的“桥梁”），从而，为进一步灵活多样的动态分析和计算机辅助处理（有针对性地重用标准化认知模型），奠定了“中文信息处理”的系统工程基础。

发现与记录

实际应用中，形式与内容，通常总是联系在一起的。凡经过形式化系统工程处理的音节或文本序列（对汉语语汇而言，就是：充分利用 GTCM的 4、5、6三个分表和 GSCM的 1-m多个分表提供的音字序列的形式信息以及 GTCM 的 0，1，2，3，4五个分表提供的形字序列的形式信息，优化中文信息处理过程的记录），在协同智能计算系统中，无论是其形式信息还是其内容信息，都将一目了然（因为经过“两表”、“三化”加“三注”加工之后的汉语及中文语汇知识信息数据处理可且易“计算、操作、重用、共享”）。用户的个性化重用，不过是该系统的标准化重用的某些具体的组合变换(分与合——有针对性地重构或重组)而已。通常有一类例外 :某个或某些特殊的用户发现了该系统未曾分析和处理过的具体组合。此时，系统将自动记录该用户或该终端的原始输入信息，并与本系统长期协作的知识工程师、领域专家以及知识产权专家一道协同对之进行复查和审核。

区分字与词的必要性

汉语的字与字组的关系（涉及“接续”问题）

区分形字与音字，是汉语形式化的一个基本问题。涉及：如何认知汉语自身发展路径与如何继承汉语研究传统的问题。对汉语“辞、块”的进一步认识和研究，主要建立在对音字

的认识和研究的基础之上。如：汉语固有的基于字（汉语“字本位”理论突出：字是汉语的基本结构单位）的“切辞块”（汉语“字本位”理论突出：基于实字的辞和在字或辞的基础之上附加虚字及虚字组的块）与“断句读”（由古代汉语延续下来，汉语“字本位”理论突出了：读的语气停顿和句的语义停顿）的困难如何解决（汉语教学和中文信息处理都关注）。

汉语与英语的结合（涉及国际“接轨”问题）

区分音字与英词，是汉语形式化的另一个基本问题。涉及：如何认知汉语融合发展路径与如何借鉴外语研究传统的问题。对英语词组或短语的进一步认识和研究，主要建立在对词的认识和研究的基础之上。自从汉语引入（外语的）词（word）概念之后，分词与标注的困难始终与中文信息处理为伴。对汉语引入（外语的）词组或短语与汉语本身的辞或块的关系的进一步认识和研究，主要建立在对字与词的语言交融现状的认识和研究的基础之上。如：引入词概念，在切辞块与断句读（对自然人）之外，又增加了分词与标注的困难（对计算机）。

7 典型示例——解析“字与字组的关系”的实例 1-2

由“一字精解”到“字字精解”的步骤 1-3，把字视为对象语言，把二字组视为元语言。

实例 1 ——中文形式信息处理（从形式方面，解析“字与字组的关系”）

示例 1：如何区分语言的“字”和文字的“字”？

步骤 1：字内信息“层解”（从复合字中分离出形字与音字的“分离手术”是难点）

对基于笔画的“层面型结构”（形字）的“层解”是基于 GTCM第 0，1，2，3，4 进阶五个分表的类及例而实现的。在计算机数据库和数据仓库中，表现为：由五组“数字（id）”与“形字（逐层分解的字符）”数据“同义并列”的五个“表”（见图 20的一览总表）。

图 20是 GTCM五分表示意图。

由图 20结合 GTCM可说明：GTCM第 0，1，2，3，4 进阶层式与计算机 FONTS（字库）以及 GBK或 Unicode的兼容关系。基于中文标准信息交换码的思路，实现形字的层面型结构化改造——使字内信息成为：可计算、可重用形式信息。

GTCM 第0（基本笔画），1，2，3（偏旁部首），4（字）进阶层式（实施例）

编号	笔画字27个	损形字28个	变形字16个	字中字162个	标准字13675个
1	一	匚	亻	一	吖
2	丨	刂	𠂉	乙	阿
3	丿	冂	冫	二	啊
4	丶	勹	口	十	锕
5	乙	冖	彡	厂	錒
6	…	…	…	…	…

图 20

由于在“GTCM”中“层面型结构”具有：“可分解性”以及“被分解后的各级部件”具有：“可计算性”，因此，从“形字”中提取必不可少的“字内信息”相当方便。由于“形字”与“音字”之间的“迭交”关系，所以，“字内信息”与“字间信息”两方面的“形式信息”提取，都是“中文信息处理”必须的。

形字与音字“迭交”原理及实例。“音字”切分为“节点”与“形字”拆分为“部件”。由“层面型结构”顶层可透视：音形“迭交”的情形。如：图 1 中“义”这个“字”，就正好位于“线串型结构”的“音字”与“层面型结构”的“形字”的“交汇处”。

步骤 2：字间信息“串解”

就“线串型结构”而论，图 1 中虽然“文本”与“本义”是两个可以直接“接续”的“字组”，而“文”、“本”、“义”则是三个“离散”的“字”，但是，它们都是字字落在“线串型结构”的“节点”上的。其它字的“字间信息”的解析与此同理。就“层面型结构”而论，图 1 中“义”这个“字”的“字内信息”，涉及一个“乂”（字中字）和一个“点”（笔画字）。其它字的“字内信息”的解析与此同理。可见：每一个“字”都有“语言”与“文字”的“双重特性”。这就是汉语“音字”与“形字”相互“迭交”的性质。

“形字”是从文字学角度得出的概念。“字形”是对“字”的“形”的研究。其特点是：从平面“方块形”结构入手。着重点在于分析“视觉信息”，表现为：基于“笔画”和“部件”或“偏旁部首”的“形”分析。

“字音”是从语音学角度得出的概念。“字音”是对“字”的“音”的研究。其特点是：

从立体“单音节”结构入手。着重点在于分析“听觉信息”，表现为：基于“音素”和“音节”及“语音语调”的“音”分析。“字音”的形态，可以“拼音化”。表现出：汉语总与拼音这根拐杖联系在一起的特点。

“字音”和“形字”是从语言学角度得出的概念。“音字”是从“音”的方面对“字”的研究。其特点是：从“线串型结构”入手。着重点在于分析“字间信息”，表现为：对“语汇”有关的“语音”、“语法”和“语义”乃至字间“语用”等“信息”的关注。“形字”是从“形”的方面对“字”的研究。其特点是：从“层面型结构”入手。着重点在于分析“字内信息”，表现为：对“语汇”有关的“文字”、“语义”乃至字内“语用”等“信息”的关注。由于“字间信息”与“字内信息”都对“义项”具有限制作用，所以，从“释义字组”的选取范围考虑，必须同时兼顾“字音”和“形字”两方面的语言信息。

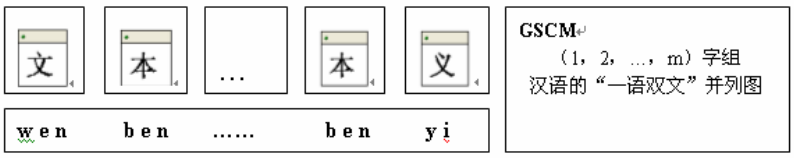
图 1 左边的“节点”切分图，展示了：“音字”外部的“连串组配”机理。“音字”特指：基于“GSCM第 1 进阶层式”的“线串型结构”。与狭义的“形字”之间是“迭交”关系。“音形字”中的“声符”可视为“音字”的特例或原始类型。

图 1 右边的“各层”透视图，展示了：“形字”内部的“分层组配”机理。狭义的“形字”特指：“迭交”于“GTCM第 4 进阶层式”的“层面型结构”。广义的“形字”特指：基于“GTCM第 0，1，2，3，4 进阶层式”的“层面型结构”。

示例 2：如何解析“文本义”这个“音字”串？

首先，以“音字”为单位，把“线串型结构”自动分解为“离散”的“音字”串，即：“文”、“本”、“义”。接着，基于“一字表”自动识别“音字”串中的“实字”与“虚字”以及“虚实两可的字”。本例断定全为“实字”。第三，基于“二字表”自动识别“音字”串中的“实字”之间“两两组合”是否符合“接续”要求。本例断定“文本”与“本义”是符合“接续”要求的“二字结构”。第四，根据“基本组字公式”分析“字间信息”。本例断定：释辞 1：“文本”=“文”+“本”=“用字”+“解字”。释辞 2：“本义”=“本”+“义”=“用字”+“解字”。第五，基于“三字表”自动识别“音字”串中的“实字”之间“三三组合”是否符合“接续”要求。本例断定“文本义”不符合“接续”要求，因此，不构成“三字结构”。这是“一字表、二字表、三字表”分别位于“GSCM第 1，2，3 进阶层式”。同理，可分析分析其它“线串型结构”。 图 21

图 21是一语双文。



由图 21可见：汉语具有“（音）字与（由音字组配的拼字）字组”和“（由字母组配的拼音）音节和音节串”两种“记音符号”。由此可见：汉语“音字”与汉语“拼音”（即：“字音”的一种表现形式）并存的所谓“一语双文”现象。图 1 与图 21结合可帮助我们更好地理解：汉语语言学中与“形字”之间“迭交”的“音字”（在图 1 中已通过“义”字的拆分与叠合的方式直观展现）和汉语语音学的“字音”之间的区别与联系，即：在图 21中上方汉语“拼字”与下方汉语“拼音”之间的“同义并列，对应转换”关系。

如果没有图 1 对“义”（形）字“层分”的直观展现，及图 21 对“义”（音）字“层分”的直观展现——“一语双文”，那么，“线串型结构”与“层面型结构”、“音字”与“形字”之间如何“迭交”通常是不容易理解的。

实例 1 和图 18、1、19 从汉语自身结构分析的角度展示“层解”限制“层面型结构”的“释义”与“串解”限制“线串型结构”的“释义”的特征。从一个侧面以“窥斑知豹”的方式说明了“字与字组的关系”。以下实例 2 将借助“两表”从英汉双语对比的角度，从另一个侧面说明“字与字组的关系”。

步骤 3：义项信息“分解”（见：典型分析 1-5）

实例 2 ——中文内容信息处理（从内容方面，解析“字与字组的关系”）

典型分析 1：“义”这个字，如果单独看，那么，它可有“本义、主义、道义、…”多个“义项”可供选择。但是，如果前面增加了“本”这个字（“节点”），那么，其“义项”选择的可能性一下子也就减小了。因为，“本义”作为“释义字组”，明确地排除了选择其它“义项”的可能性。再延长“字组”的长度，还可有“易经本义、圣经本义、他的本义、你的本义、…”多个进一步的“义项”（注：不过这里的“义项”已是对“本义”这个“字组”而言了）可供选择。一旦前面增加了“你的”这个“字组”，其“义项”选择的可能性立即也就减小了。因为，“你的本义”作为“释义字组”排除了其它进一步“选择”的可能性。…

典型分析 2：仅就形式方面而论，“字”这个字，如单独看，可有多个“义项”可供选择。如何才能以最简洁的方式消除这里的歧义（即：二歧性）呢？根据“基本组字公式”，只须在其前面增加一个“用字”即可立即明确地消除“字”这个“解字”的“二歧性”——如果要求在“音”与“形”两个“用字”之间二选一。这里，“音字”或“形字”就是“释辞”。“用字”的功用由此可见一斑。这样，我们一旦明确地说：汉语“字本位”理论所说的“字”是“音字”而不是“形字”，那么，人们就可以立即断定：汉语“字本位”理论所说的“字”属于语言学的研究范围。因为只有“形字”才属于文字学的研究范围。

典型分析 3：把“典型分析 2”的研究再向前推进一步，在“释辞”中扩大“字”这个字的义项选择范围，即：扩大到“形字、音字、实字、虚字、用字”。 图 22

此表着重说明：汉语的“形字、音字、实字、虚字、用字”之间的相互关系。

语用学	用	字	释义	互用信息	释义用字	组字释义	语汇 去限制而释义
语法学	虚	字	关系	字间信息	测序定位	组语成句	语汇 靠关系释义
语义学	实	字	概念	对象 信息	选域定向	组字成语	语汇 被限制而释义
语音学	音	字	表音	单音节	听音定调	线串型结构	语汇 音形“迭交”
			拼音	字外信息			拼音是拼字的辅助
文字学	形	字	拼形	字内信息	偏旁部首	层面型结构	语汇 靠符号释义
符号学			符号	笔画 信息	基本笔画		拼形是拼字的基础

从语言“音义结合”的“形式化”方面来看，在“形字、音字、实字、虚字、用字”中，“音字”位于“线串型结构”与“层面型结构”的“迭交”结合部，具有独一无二的中枢地位，是汉语“拼字成为语句”（字组皆由拼字音节的音字构成）的基本特征。故汉语当属“拼字”语言。

图 22是“形、音、实、虚、用”关系示意图。

无论是与各国语言相比较，还是与语言学的各种理论相比较，都有证据说明：在“形字、音字、实字、虚字、用字”中，汉语的“音字”是独一无二的。如此显著的特征，为什么会被学界（长期地）视而不见？难道“音字”存在的现象不是事实吗？还是其中另有原因？否则怎么会长期存在“一叶障目”的情况呢？

证据：（从理论上讲）造成这种“一叶障目”的主要原因可能是：古代汉语研究缺乏科学的语音学指导，而现代汉语研究又因为引入科学的语音学的同时实行了“汉语拼音方案”（之后又产生了所谓“一语双文”的情况——这是相应的实情）。（从实际上看）“音字”存在的现象，在汉语中是一个事实。如：“诗经、楚辞、汉赋、乐府、唐诗、宋词、元曲”等经典的存在，都说明：在古代汉语中“音字”的特点，事实上是被认可的。在拼音体系还没有引入中国之前，不仅古代汉语就是现代汉语形成初期的白话文的流传过程中，汉语“音字”的特点，在事实上也是被认可的。“音韵、训诂”之学也记载并保留了“事实认可”。

在拼音体系引入中国之后，加速了白话文和现代汉语的普及进程，特别是汉语拼音体系（如：“汉语拼音方案”推行的结果）建立之后，随着普通话的推广，汉语出现了“一语双文”。于是，汉语在字形与字音（即：拼音）之间“分工”过程中，人们有意或无意地选择：用“字形与拼音之间容易区分的明确形式——字音”取代了“字形与拼音之间难以区分的不明确形式——音字”。这样，在汉语“一语双文”普及进程中“与其说被取代不如说被掩盖”的正是：（与“形字”同形的）“音字”。

典型分析 4：把“典型分析 1”的研究也向前推进一步，在“释辞”中扩大“义”这个

字的义项选择范围。即：调整“解字”与“用字”关系。

图 23是“义项”与“释义字组”的（直呈）关系示意图。　　图 23

由图 23可见：“义”这个“解字”的义项，是通过具体的“用例”（即：等价于包含“解字”的“释义字组”）直接呈现的。这说明：“字的义项”与“释义字组”之间的直接关系，可通过“线性组配”限制“释义”的直接呈现方法（即：“左右限制法”），围绕“核心字”展开。根据“基本组字公式”，“本义、主义、道义、…”多个“释辞”直接呈现的“义项”都是由“本、主、道、…”等“用字”的限制功能而发挥“消歧”作用的。其它可类推的部分省略。

字与字组的关系（涉及：内容与形式，即：字的义项解释字组化）示意图

此表是对“义”这个字的“义项”的解释的“字组”的抽样示例

…	2	1	2	3	4	5	GSCM 序号
		义					
	本义						
	意义						
	主义						
	道义		义举		义不容辞	义正词严地	
			义务	义务工	义务劳动	义务劳动者	
	…	…	…	…	…	…	

所有字的义项的每一个用例均可视为等价于包含该字的各个具体的字组。

典型分析 5：把“典型分析 1与典型分析 4”的研究再向前推进一步，把“释辞”由“直接呈现”扩大到“间接呈现”。

如前所述，义项呈现字组化，包括：直接呈现与间接呈现或信息标注（即：“三注”）。细分的（同义并列的双语）“义项”说明，相当于：细分的（同义并列的双语）“释义字组”以及“常识”和“知识”等“领域”的“标注字组”的说明。

释义字组直接呈现义项（字与字组的关系）				间接呈现的“常识和知识“涉及：释义“字组、句子、段落、篇章、…”							
				语言文字信息标注		通用常识信息标注		专用知识信息标注	
…	2	1	2	（涉及若干“列”标注---释义字组）		（涉及若干“列”标注---释义字组）		（涉及若干“列”标注---释义字组）	
		义							
	本义			Original	…	语义常识	…	语义学领域	…
	意义			Meaning	…	哲学常识	…	语言哲学领域	…
	主义			-ism	…	政治常识	…	政治学领域	…
	道义			Moral and justice	…	道德常识	…	道德学领域	…
			义务	Incumbency	…	法律常识	…	法学领域	…

图 24是“字的义项”与“释义字组”以及“标注字组”（直接和间接呈现）的关系示意图。

由图 24可见：通过汉语直接呈现的“义”这个字的义项的“用例”不仅可与英语的对应词语之间实现双语的同义并列（由此也发现汉语与英语之间的显著区别——“对译”的语言单位并不一致），而且，还可通过汉语的“释义字组、句子、…”的方式进行多角度或多领域地“间接呈现”（即：“三注”，这里仅限于“释义字组”）。“三注”，是通过多个领域的标注信息“立体选配”，达到进一步限制“释义”范围的“间接呈现法”（即：“行列限制法”），围绕“（领域）参照系”展开。

8形式化探新简议

“形式化”通常是就“形式语言、程序语言、人工语言”而言。“美国标准信息交换码”（ASCII）是这种“形式化”的基础。与“英文信息处理”比较而论，“中文信息处理”至今没有自己独立的基础。“统一编码”（Unicode）虽然提供了“国际标准”，但是，仍不能改变汉语与英语在此基础方面的根本差距。有一个办法可消除这个差距。这就是建立既能与ASCII和 Unicode兼容，又能与 ASCII平级的“终极标准信息交换码”（Z-ASCII和Z-Unicode）。本文的“字内信息”处理，有利于这个问题的解决。

“字内信息”由“GTCM的 0-4分表”处理。如果这个工作得到相应的资金支持，我们就可以早日开发出基于 Z-ASCII和 Z-Unicode的中文输出输入系统（Z-BIOS）和小字符集中文字库（Z-FONTS）。Z-BIOS与现有的英语 BIOS兼容且平级从而可用汉语直接控制。Z-FONTS与现有的大字符集汉语 FONTS兼容且与小字符集的拼音字库平级从而可用汉语直接控制。如果这个工作得到普及，就可开发出能在底层用汉语直接表达的软件开发平台。

在此基础上“字间信息”由“GTCM的 4-6分表”构成“汉语字组粗分模型”或由“GSCM的 1，2，3，…，m分表”构成“汉语字组细分模型”处理。

完成上述两步，才可说“中文信息处理”真正上了一个大台阶。因语言处理与知识处理相辅相成，所以，必须继续前进，完成“字外信息”由“GTCM的5-12分表”处理的过程。

完成上述三步，才可说“中文信息处理”真正融入了“自然语言处理”的大家族。

如果知识处理不能上一个大台阶，那么，语言处理也难以跟上国际科技前沿的发展。

由于现代知识信息数据的创新部分大部分是以英语公开的，所以，除了解决汉语字与字、字与字组、字组与字组的语法接续问题外，还必须关注汉语与英语的国际接轨问题。

因此，汉语的字与英语的词之间的“中介”——“释义字组”与“释义词组”（均由“GTCM的5-6分表”进行“形式化”处理），也就成了本文关心的一个重要部分。

“用汉语思考与表达”与“用英语思考与表达”能否地位平等？关键在于“对象、概念、符号、关系”的“释义字组”与“释义词组”之间，能否成体系地掌握到位？

就字与字组的关系而论，如果从语言事实中发现的迭交原理、等价原理、释辞公式和语块方阵能为完成上述三步提供可计算、可操作、可重用、可共享的路径，那么，不仅汉语字本位理论体系可完善，而且，其优越性也将举世公认。

那时，基于汉语且兼容英语的高性能计算机以及基于Z-SCII和Z-FONTS的中文操作系统（Z-OS）也才有可能出现。Z-OS与英文操作系统兼容且平级从而用中文直接解释的程序语言控制，区别于基于英文操作系统的“汉化”或翻译的中文操作系统。

在形式上，本文的模型建立在数字计算机及其关系数据库和数据仓库的基础上；在内容上，是基于“相对完全归纳”的“语言事实”（如：“现代汉语词典”）集合。经过“三化”处理的“模型”是“标准化与个性化结合”的“理想化认知模型”，是当代逻辑学、数学、计算机科学、认知科学乃至人工智能技术与汉语“字本位”理论的有益结合。

结语

以上主要介绍了汉语及中文信息处理的形式化体系，涉及：字与字组的关系数据原理。问题的提出源于理论融智学对“语义三棱”（模型）的解析和工程融智学对“意义=意+义”（字符串公式）的解析。问题的解决得益于“字本位（汉语的基本结构单位）”的启示和字的“迭交”现象的形式化虚拟描述。分析过程涉及“字内、字间、字外”信息处理三个步骤。由此提炼出“字本位与中文信息处理”的全面形式化方法，其基础是：字的“迭交”原理。

基于笔画的形字，字字可重构/再造（根据笔画表元素顺序编号id所组成的各形字的笔画排序结构编号ip可随时随地调用/重用形字以便进行有针对性地各种组合变换）；基于音节的音字，字字可重用/再造（根据音字表元素顺序编号id所组成的各音字/音节字组的组成排序结构编号ip可随时随地调用/重用音字以便进行有针对性地各种组合变换）。借助“文本结构控制模型（STCM）”和“音节结构控制模型（SSCM）”可实施“形字”与“音字”的“分体手术”。这样，既可在实践中有针对性地调用/重用“形字”与“音字”，有可在理论上有力地论证汉语“字本位”理论的基本原理“字与字组的关系”（含：字与字组的科学定义——不仅可定义各自唯一的“类”，而且还可在相对完全归纳的范围之内枚举各自的“例”）。

综上所述，本文不仅直接证明了“三化”（汉语及中文的形式化）的必要性、重要性和可行性，而且，还间接概述了“三注”（知识处理）的必要性、重要性和可行性乃至紧迫性。读者可以经验主义（Empiricism）、理性主义（Rationalism）和怀疑主义（Skepticism）三种观点，检验本文的方法及结果的科学性（可重复性或可计算性），质疑任何不可验证之处。

值得进一步研究和思考的几个问题。如：现有的字典和词典没有指出这样的问题，即：字字有（语义）分歧，处处有（语义）陷阱。这似乎只有在本文所述标准平台及其相对完全归纳的“两典一册”完善且普及的网络与计算机辅助的环境或条件下，才可能较好地解决或妥善处理。又如：本研究发现：汉语及中文的语法与英语及英文的语法有一个很大的不同，即：前者以字法和章法这两极的系统发展为特点；后者以构词法、词法和句法这三级的系统发展为特点。再如：笔者发现：古代汉语有相当发达的字法和章法，而现代汉语似乎没领悟

中文语法的这个特点。

参考文献

李谷城等译：现代语言学（乔姆斯基革命的结果）[M]外语教学与研究出版社 1-320 页 1983

方　立：美国理论语言学研究[M]北京语言学院出版社 1-240页 1993

喻云根：英汉对比语言学[M]北京工业大学出版社 69-99页 1994

石　锋：汉语研究在海外[M]123-188页，北京语言学院出版社 1995

张志公：汉语简论[A]汉语辞章学论集[C]人民教育出版社 1996

刘叔新：词语强制搭配的语义关系类别及其性质[A]南开大学语言学论辑[C]北京语言学院出版社 1996

徐通锵：语言论--语义型语言的结构原理和研究方法[M]东北师范大学出版社 1-442页 1997

黄增阳：HNC（概念层次网络）理论——计算机理解自然语言的新思路[M]清华大学出版社 1998

邹晓辉：融智学原创文集[C] www.survivor99.com/pscience/mysite%201-20/index.htm 2000-2005

北京大学计算语言学研究所：计算语言学文集（第 4 集）[C] 1-254页 2000

徐通锵：基础语言学教程[M],北京大学出版社 19-36页，178-237页 2001

鲁　川：汉语语法的意合网络[M]1-277页，商务印书馆，2001

施伯乐等译：数据库处理——基础、设计与实现[M]电子工业出版社 170-246，334-489页 2001

康博创作室：SQL Server 2000 数据仓库设计和使用指南[M]清华大学出版社 2001

冯志伟：发挥汉语拼音在信息时代的作用[A] 语文现代化论文集[C]商务印书馆 41-44页 2002

黄河燕主编：《机器翻译研究进展[C]电子工业出版社 1-282页 2002

苏培成等：语文现代化论文集[C]商务印书馆 1-364页 2002

张学文：组成论[M]中国科学技术大学出版社 44-56页，246-252页 2003

http://www.survivor99.com/pscience/　　http://survivor99.com/entropy/