EasyChair Preprint

№ 15754

Original Collection on Smart-System Studied

Xiaohui Zou

January 24, 2025

融智学原创文集

作者：（珠海）邹晓辉（Zou Xiao Hui）

痴人视机会而不见，
常人只知等待机会，
智者积极寻找机会，
天才努力创造机会。

1989 年春于深圳

（本文集的文章均公开发表于 2000-2005 期间）
均被“前沿科学”和“信息学报”网络学术期刊连载

目　　录

《融智学原创论文集》第一版（2000-2005期间公开发表的20篇原作）
《融智学原创论文集》第二版（附录部分:增加部分专家评语和文章）

前　言

融智学是一门研究自然人与计算机“合理分工、优势互补，高度协作、优化互动”的原理、方法及实例的学问。融智体系形成经历了三个发展阶段（其间伴随着相应的理论思考和社会实践）：

第一阶段是从1976（产生“集人类知识之大成”的构想）-1992（形成“智慧融通”或“融通智慧”的概念），标志是提出“信息基本定律（假说）”，即：“同义并列，对应转换”法则。1980-1981笔者向贵州大学物理系现代化实验室刘汉云主任、人才学创始人雷祯孝、《自然信息》张良主编等学者做过介绍。1987发表“企业招标投标中法律顾问的作用”（优秀毕业论文）。同年， 撰写了“贵州省七五社会科学项目人才课题招标投标的全套标书”。1989依托深圳“图书馆自动化集成系统”试办“全方位”咨询与培训服务。1991公开“心理学科学体系探新”（中 国心理学会基本理论专业委员会心理学基本理论年会论文）、“终身教育学纲要”和“考试心理学——学习、复习、考试与应变的原理”（纲要）。1992公开“社会主义公有制与公司（法人）制度结合是我国经济法建设的突破口”（十三省、市、自治区第八次经济法学术研讨会论文）。

第二阶段是从1993（“一种智能通信子母机”的发明专利公开）-1999（完成“一种知识信息数据处理方法及产品”的构想），标志是开办“企业知识产权战略”专栏。这期间亲自发明和指导发明的多项新技术分别获“中国专利技术博览会”和“国际发明博览会”金奖、银奖和铜奖多枚（1994-1997）。同期，还尝试了系列“融智与融资” 项目的策划和组织实施。

第三阶段是从2000（“一种知识信息数据处理方法及产品”的发明专利公开）-2005（编著的“融智学原创文集”集中发布），标志是形成（基于网络和计算机辅助的）“融智学”（三部曲），即：着重基础研究的理论融智学——强调“语义、信息与智”的统一理论框架；着重间接计算的工程融智学——强调“知识信息数据处理”的间接形式化；着重间接融资的应用融智学——强调“产、学、研、用、算”一体化管理。“字本位与中文信息处理的基础，合作型生产式教学法，融智与融资”是涉及多个产业链及产业群的三个典型的融智实例。

“融智学原创文集”主要是为关心“网络和计算机系统与用户（自然人）如何互助”和“融智与融资如何互补”两个研究方向及实际应用领域感兴趣的读者而编著的。

由于融智学是与基础语言学、计算机科学、认知科学、机器翻译学、计算语言学、人工智能学、一般信息学、知识管理学、知识经济学、网络和计算机辅助教学方法和知识产权法学等多学科交叉的一门新兴学科，因此，也适合这些领域的有关学者和一般读者。

北京大学中文系“全国普通高等学校人文社会科学研究十五规划纲要”语言学咨询组负责人徐通锵教授（2001，9，11）来信说：义、文、物、意概念的提出，我觉得是有价值的，但如何阐述？抓住什么样的核心？需要深入推敲。这四个概念的关系，据我的理解，“义”应是客观存在的事物的结构机理，或者说是客观规律，其运转规律不以人的主观意志为转移；“意”是主观对“义”的认识或理解，“文”与“物”只是这种认识外化的表现形式。区分“义”与“意”是很必要的，现代语言学也已意识到这种区分的必要，其具体的表现形式就是注重功能的研究。如何将这种区分进行理论上的阐述，还需学界的努力。

中国人民解放军洛阳外国语学院计算语言学研究室易绵竹教授（2001-09-25）来信说：融智学新范式提炼出协同智能主体的概念体系具有原创性，想必对自然语言语义信息的处理将引发一场革命。

清华大学智能技术与系统国家重点实验室苑春法教授（2003.1.3）来信说：“谢谢你在清华的讲座。由于时间关系，不能长谈。仅仅从几个小时的讨论交流中对你理论全貌尚未能得到一个清晰的了解。从交谈中，我认识到你的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。但是同时我也认为这13张表的构建是一件消耗大量人力物力的工作。因为仅仅一个汉语的树库的建立就是一件浩繁的工作，至今尚未完成；而它仅仅是你的数据库中的一部分。所以我建议在经过充分酝酿和充分的人力财力准备的基础上再启动这件事。”

教育部国家语委语言文字应用研究所鲁川教授（中文信息学会首届计算语言学专业委员会主任）来信说：这13张表的构建充分体现出你（指：笔者）能站在一个较高的起点上善于集中现有各家学派的优点。

……

目前《融智学原创文集》主要汇集了融智学作者2000-2005期间分别在学术期刊和学术会议上发表的文章。虽说作者现在（2006）的认识已比过去三个时期的任何一个阶段（1976-1987，1988-1992；1993-1996，1997-2009；2000-2002，2003-2005）深刻且系统化了许多（注：这将在修订更新的部分陆续追加补充）。但考虑到原创文集独特的学术交流价值，例如：其中保留着对科学技术乃至应用的原创成果涉及与基础语言学、计算机科学、认知科学、机器翻译学、计算语言学、人工智能学、一般信息学、知识管理学、知识经济学、网络和计算机辅助教学方法学和国际知识产权法学等多学科发生交叉的一系列问题，尤其是涉及一门新兴学科如何处理或表述与众多相关学科之间关系的各种尝试或探讨，对从事跨学科知识创新的研究人员来说，都将会有借鉴、启迪或警示作用。

同时，也有必要保留原创成果初创时期的基本风貌，而这在研究生的教科书中也是无法看到的。

据笔者与上述有关多个领域的学科带头人直接交流的经验、感受和体会来看，本文集所具有的创新知识点是有相当学术价值的，有些还有广泛的社会经济实用价值。

欢迎广大读者多提出宝贵意见！

希望就新兴学科如何处理与周边交叉学科关系的问题，与读者进行有益的科学探讨！

尤其希望能与对“协同智能计算系统”和“融智与融资”双重实践的读者交换意见！

作者 邹晓辉2006-3-25于珠海恒美花园

注：作为融智学导论的《字本位与中文信息处理的基础》和作为融智学三部曲的《理论融智学》、《工程融智学》与《应用融智学》几部专著除由《融智学原创论文集》的有关部分提炼、升华和发展而来之外，还吸收了“融智学精华介绍”、“融智学纲要”和“融智学术语表”的内容。而作为精品课程的《融智学简明教程》则有待笔者及其弟子们在教学及其组织管理实践中不断总结或提炼其表述形式。


国家知识产权局
中国专利信息中心
Sponsored by
China Patent Information Center,SIPO
中国专利技术博览会
CHINA FAIR OF INVENTIONS AND NEW TECHNOLOGIES
SINOPAT
金奖
Gold medal
证 书
Certificate
获奖项目
in recognition of the display of
一种知识信息数据处理方法及产品
获奖单位
presented to
邹晓辉
编号
Serial number
2001T2607.1
颁奖日期
Awarding date
2001.6.12
第八届中国专利技术博览会
组织委员会
2001

一种知识信息数据处理方法及产品[1]

本发明涉及人工智能、计算机和通信技术的交叉综合领域，属于一种语义信息及真实文本的数字处理技术，进一步是一种知识信息数据处理方法及产品。

现在，虽然在知识工程、信息技术和数据或数字技术等方面有很大的发展，但是，由于人类关于语义信息的定性分析、定量分析和结构分析长期未能获得实质性的重大突破，因此现有技术至今仍无法解决以下一系列知识信息数据处理的难题。

数据或数字处理作为现有技术的核心，对形式信息的量化处理虽然十分有效，但是，对语义信息的量化处理却难以直接派上用场。因此，各种各样的中介技术方法及产品，便介入并参与了对知识信息及真实文本的量化处理，尽管其效果大打折扣。在受限范围现有技术虽有较大进展，然而，在非受限范围却一筹莫展。因为，现有技术只能对受限范围的知识信息及真实文本进行局部的量化处理，所以，不能从根本上解决冗杂文本、非标形式、垃圾信息、知识爆炸和怪圈悖论等难题。

目前，还没有能够一揽子解决上述难题的技术公开。可以说，就解决上述难题而论，不仅现有技术与相关技术的整合优势没有形成，相反，各自为政、一盘散沙的劣势却无处不在。根本谈不上，在人工智能与人类智能、电脑与人脑、电信网与神经网之间实现协同效能或优势互补。协同智能的发展在知识信息数据处理领域受到了现有技术的瓶颈制约。至今为止，尚无与本发明相同或相似的方法及产品问世。

以下进一步举例说明现有技术的缺陷或不足。

以知识信息数据处理为例，无论是直接的模数转换还是间接的编码转换，要么由于真实文本未经提炼就直接转换为数字符号（例如：数字图书馆技术），根本谈不上对知识信息的量化处理；要么由于真实文本与数字符号之间的中介程序太多太杂（例如：计算语言学及其应用技术），根本无法形成对知识信息及真实文本进行量化处理的统一标准，现有技术只能是一盘散沙、各自为政。

进一步，以编码转换为例（涉及语言形式和知识内容的处理），一方面，现有的语言信息处理系统（包括机器翻译、自动识别以及广义文本通译等），还存在各种语言形式之间的通译和同种语言的非标形式的识别等一系列技术瓶颈；另一方面，现有的语义信息处理系统（包括机器分类、自动浏览以及知识基因提取等），也还存在诸如冗杂文本、垃圾信息、指数爆炸和怪圈悖论等一系列难题。现有软件工程及知识工程体系还缺乏从受限范围到非受限范围的转换机制，缺乏总体标准。

再进一步，以金融监管（涉及证券、期货、外汇交易各方对价格信息的量化处理）为例，造成失误或失败的原因虽很多，但根本原因是：人们还没有找到能够对各种价格变化所包含的激励或预警信息进行及时而透彻分析的有效方法（包括定性分析、定量分析和结构分析）。人们普遍认识不到：金融的本质是智融，缺钱的实质是缺智。现有的专家系统，无论是终端还是网络形式，都不具备协同量化处理知识信息及真实文本的功能；现行科教体系熏陶或培养出来的专家，即使是最有知识的专家，也只能处理非常有限的知识信息及真实文本，而且，还常常伴随着以偏概全的议论或见解。由此可见，人工智能和人类智能，都面临着协同智能的严峻挑战。

推而论之，在不同方面、不同阶段、不同层次和不同系统，都能轻而易举地发现现有的教育、科技、经济、政治、外交、军事、法律、医疗卫生和日常生活等几乎任何一个领域都面临上述难题和挑战。在如今这样一个充满竞争且快速变化的竞智时代，试想一想：如果一个人、一个单位、一个国家，总是处于冗杂文本、垃圾信息、指数爆炸和怪圈悖论的包围之中，既搞不清楚自己所收到的这么多知识信息的本质含义，又不知道如何准确无误地发出自己应该发出的有的放矢的知识信息，也不明白自己所处的内、外、大、小环境中实际存在的各种重要的知识信息（包括正面的激励信息和反面的预警

---

[1] 2000 年 11 月 29 日国家知识产权局专利局《发明公报》第 16 卷 48 期
2000 年 12 月（徐辉任主编时）公布于　　系统科学之窗 论文专区
2002.04 (张学文研究员任主编时)公布于　　潜科学（学术期刊）

信息）的基本内涵，这时，如果这个人、这个单位、这个国家又处于不利的竞争地位，甚至面临危机，那么，其处境或命运将会是怎么样的一种状态或过程呢？其结果是可想而知的。如果一个系统，包括智能网络及终端乃至独立的机器人，总是被非标形式或垃圾信息所阻止或干扰而无法正确应变，那么，其结果也是可想而知的。

现有技术的发展还受到当前的语义学理论和语义信息处理理论的发展制约。产业界对知识信息及真实文本的量化处理技术的效率，之所以如此低下，这与学术界对语义信息的本质认识或阐述不清，是息息相关的。不仅仅是普通人把形式信息与语义信息混为一谈，而且绝大多专家也都常常把衍生形式与本真信息混为一谈。至今为止，还没有人明确地区分本真信息、形象符号、载体载能和意向意识，并且以此作为重构人类整个知识概念体系的基础框架。虽然个人计算机和互联网技术接二连三地推动着信息技术不断地向前发展，但是，语义信息的本质究竟是什么？至今仍未见有令人信服的报道公开。对语义信息的定性和定量表述，还无法令人满意。

人类的整个知识概念体系，特别是语义信息理论体系，至今仍然是一个大杂烩，其根基是不牢固的。例如：以唯物论、唯心论和形式论三个基石为不同支点而形成的各种基本观点，以及在它们的基础之上构建的人类知识概念体系及其各个分支理论，都忽略了本真信息的根本地位，甚至把本真信息与载能载体、意向意识、形象符号等衍生形式之间的基本关系本末倒置。在这种情况下，要发明量化处理知识信息及真实文本的有积极效果的新技术，首先就必须要找到能重构人类整个知识概念体系的新理论，否则，就根本不可能超越现有科技框架而获得真正的重大突破。

本发明的目的是提供一种知识信息数据处理方法及产品，包括：文化基因工程方法以及相应产品的生产及使用方法；全域数码定位系统及其派生产品。通过对全域基因文本元素的完全归纳和对已知域及目标域基因文本组合的相对完全归纳，解决知识信息的计量及测度的难题，促使人工智能与人类智能优势互补，形成效率更高的协同智能，促成形式信息革命向语义信息革命的时代飞跃。

如果说信息论创始人仙农提出了形式信息数据处理技术及标准，那么，本发明的目的就是提出语义信息数据处理技术及标准；如果说全球定位系统（GPS）、柔性加工系统（FMS）和横断扫描仪（CT）是针对载体载能进行的全球数字定位、柔性加工和横断扫描，那么，本发明的全域数码化的网络、出版物和终端就是对知识信息进行的全域数码定位、柔性加工和横断扫描；如果说人工智能、电脑和电信网在形式信息数据处理技术的支持下获得了极大的发展，那么，本发明就是要使以协同网络、协同终端和出版物为特点的协同智能在语义信息数据处理技术的支持下获得前所未有的发展，早日迎来智能主体进化发展的新阶段——协同智能时代。

本发明的具体任务是：一、提供文化基因工程方法即全域数码定位方法；二、提供全域数码定位系统，产品形式包括：1、纯形式的全域数码化网络，即知识信息数据处理领域的“GPS”；2、纯文本的全域数码化出版物，即知识信息数据处理领域的“FMS”；3、纯数码的全域数码化终端，即知识信息数据处理领域的“CT”。

本发明依据融智概念体系和信息基本定律而设计并实施。

所谓融智概念体系（注：涉及对人类现有的整个知识概念体系的量化重构），是指以义、文、物、意为基础而构成的协同智能主体的知识概念体系。其量化形式采用能够通过复数域及复平面乃至曲面出入多元数系及多维空间的四元数形式。义，指本真信息；文，指符号形象；物，指载体载能；意，指意识意向。文、物、意统称本真信息的广义文本。形式信息涉及文、物，其中，文，包括图、文、数（注：数字信息技术属于此范围）、表、音、像等形式，物，包括立体、活体等形式；语义信息涉及义、意（注：现有的意义理论及语义信息技术没有明确区分义与意）。

以曲、棋、语言为例，对上述概念及原理的基本含义说明如下：曲、棋、语言的机理（含：法则），是本真信息，即义；展示其机理的文化形式，如：乐谱、棋谱、文字或字母或动作等，是符号形象，即文；展示其机理的物化形式，如：琴、棋、传感器官（含使用过程）等，是载体载能，即物；演奏者、下棋的人、智能主体的选择（包括以虚拟或实体的形式体现的意），是意识意向，即意。

所谓协同智能主体，是由人工智能和人类智能构成的新一代智能主体。其功能是对体现义的基因

文本元素及组合（包括程序、结构以及框架等）进行完全归纳及相对完全归纳，以及，由此发展的知识信息数据处理（包括广义及狭义的处理）能力，具有专家系统与专家群体的整合或综合优势。

所谓信息基本定律，即：本真信息，唯一守恒；基因文本，对应转换；基因通式，序趣简美；特式特例，非非各平（即：非对称、非同步、各自平衡）。

唯一守恒法则，体现本真信息的唯一性和守恒性；对应转换法则，体现图、文、数、表、音、像等本真信息或基因信息的多元基因文本（注：因为文、物、意，都是义的展示，具体表现为多元基因文本元素及其派生的各种基因文本组合）只要同义即可并列；序趣简美法则，体现在文化基因通式及基因文本通式之中；非非各平状态极其转换或变化过程，是指各种各样的特式或特例与通式相比较，具有空间上的非对称性和时间上的非同步性，以及特式特例各自平衡及趋动的特性。

本发明的目的是通过下述方案实现的。

基本方法：

一种知识信息数据处理方法，是对语义信息及真实文本进行定性、定量及结构分析的文化基因工程方法，其特征是：从广义真实文本中提取、剪接或重组文化基因，步骤是：以完全归纳的全域为基准参照系对基因文本元素的复用次数或复制件数进行自动统计，以相对完全归纳的已知域及目标域为应对参照系对基因文本组合的复用次数或复制件数进行自动计量，其中，基因文本元素及组合均采用码式并列的形式，包括式隐码显或以码代式与码隐式显或以式代码等特殊形式，即在全域数码文本体系与多元基因文本体系之间建立对应转换关系。

相应的基本产品的生产方法：

一种知识信息数据处理产品的生产方法，是以码代式进行知识信息数据处理的方法，其特征是：选择纯数字形式和纯载体形式并使之相结合构成全域数码定位系统，步骤是：采用（a+bi&…）的具体数字作为指代基因文本元素的标准代码；采用卡、表、库、网、端的具体载体作为承载基因文本元素及基因文本组合的标准载体；以全域标准代码构成基准参照系；以已知域及目标域标准代码组合构成应对参照系。

相应的基本产品的使用方法，同时，也是派生产品的生产及使用方法：

一种知识信息数据处理产品的使用方法，是以式代码进行知识信息数据处理的方法，其特征是：使用并依托全域数码化网络的基准参照系及应对参照系，步骤是：通过码式并列，把广义真实文本中的基因文本元素及组合全域数码化，构成实用的多元基因文本，即全域数码化出版物；通过码隐式显以及多元基因文本分析，支持实施知识系统工程。

进一步的派生产品的生产及使用方法：

一种知识信息数据处理产品的使用方法，是以码代式进行知识信息数据处理，步骤是：使用并依托全域数码化网络及全域数码化出版物，通过式隐码显，构成数字的一元基因文本，即全域数码化终端，支持实施网络及网际知识产权监管和端到端的默契通信。

相应的基本产品：

一种知识信息数据处理产品，是由码、卡、表、库、网、端构成的全域数码定位系统，包括纯数字形式和纯载体形式，即全域数码化网络，其特征是：纯数字形式，采用（a+bi&…）的形式语言编制源程序，采用（0&1）的形式语言编制目标程序，（a+bi&…）与（0&1）的交集即编译程序；纯载体形式，采用卡、表、库、网、端的具体形式承载码。

相应的派生产品：

一种知识信息数据处理产品，其特征是：由全域数码化网络支持的全域数码化出版物，包括已知域集大成共享基因文本和目标域集小成独享基因文本，含公开或保密、标准化或个性化、通用或专用的多元基因文本出版物。

进一步的派生产品：

一种知识信息数据处理产品，其特征是：由全域数码化网络及全域数码化出版物支持的全域数码化终端，含交换机及服务器和用户计算机及其它终端装置或载体的一元基因文本终端。

详细步骤进一步综合说明如下：

1、确定基准参照系（化无限为有限）和应对参照系（变抽象为具体），步骤是：

第一，确定基准参照系：逐方面、逐阶段、逐层次、逐系统地分解目标域广义真实文本的元素组合，直到确定基本元素及其构成的全域，并以此作为反过来判定元素组合的基准参照系，第二，确定应对参照系：从已知域的元素组合中选择具体的进化形式，涉及基本元素的时间序列和空间结构乃至元素组合的整合架构或框架，作为进一步判定复用的元素组合的具有明确针对性的应对参照系。

对目标基本元素在基准参照系中的波、粒、场特性进行全域定位，对目标元素组合在应对参照系中的时空变换特性进行自动对应，判定本真信息元素及组合与相应的符号形象、载体载能、意向意识的基因文本元素及组合之间的相互关系，进而识别并理解广义真实文本所表达的本真信息以及主体的意向意识。

基因信息元素，通常以符号形象或载体载能的某一种或多种形式体现，例如：26个英语字母（文字基因文本元素），又如：4个核苷酸碱基（生物基因文本元素）。

全域，由所有的基本元素构成，是对一定时空条件下基本元素构成的元素组合进行定位分析的基准参照系。例如：通过统计字母在字母组合中的序位和统计字母以字母组合被复用的次数，可以区分字母组合在相应的时空条件下的词素、词、词组、句、段、篇、章等具体进化形态。同理可以区分由核苷酸碱基构成的密码子、氨基酸、蛋白质、细胞、微生物、植物、动物、高等动物、人类等具体进化形态。

2、以本真逻辑为序（化无序为有序），步骤是：

第一，采用实数（a）指代本真信息，第二，采用虚数（bi&…）标识广义文本（涉及：符号形象，包括图、文、数、表、音、像等多元基因文本类型），第三，在基本元素的层次上确立体现本真逻辑的文化基因通式（a + bi&…），使其它所有的符号形象及载体载能乃至意向意识的各种进化形式体系统统简化成为文化基因通式的某一类型的广义基因文本组合。

（a + bi&…）的代数形式是多元数系、几何形式是曲面空间、分析工具是曲面坐标，四元数及复数都是它的特例形式。

基因文本的全域数码化，相当于给每一个基因文本元素镶嵌了一个唯一的“体表特征”——全域数码，基因文本组合随之也由全域数码组给大上了相应的烙印。

由此可见，全域数码体系是统一多元基因文本的标准参照系。这就是对广义真实文本进行协同操作的依据。有了它，不仅人类智能主体对各个具体的形式体系进行协同操作成为可能，而且，人类智能主体与人工智能主体协同运行也成为可能。

3、明确代码与载体（变失控为可控），步骤是：

第一，给多元基因文本元素镶嵌全域数码，并设置相应的单元载体，即：由式到码（文）到卡（物），进而，随着时空变换，由码（全域数码）到组（全域数码组），由卡到表。这就是对广义真实文本进行去冗存要处理（含广义和狭义的处理）的定位工具。

第二，明确产品形式：1、只处理数字化的全域数码的网络（包括全域数码化出版物和终端），2、用于自然人识别的全域数码化出版物（包括：公开和保密、专用和通用、共性化和个性化的出版物，3 、用于机器识别的全域数码化终端。

图1是本发明的数学模型原理图。

图 1

图2是本发明的物理模型原理图。

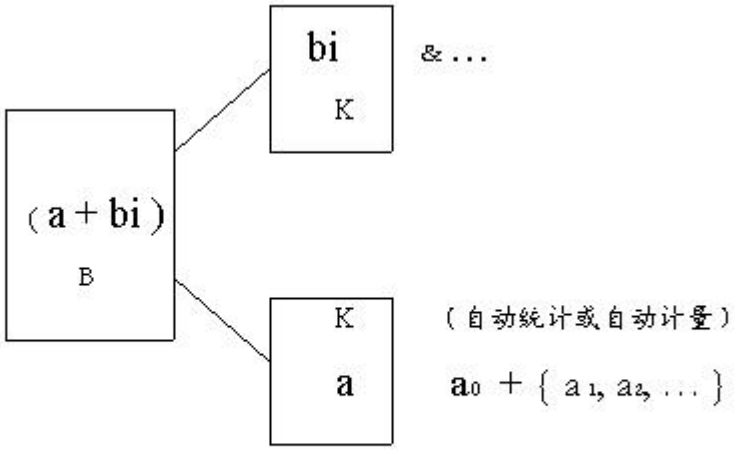


图 2

图3是本发明的技术性能一览表。

算:	DXsjk	-	GXsjk	机器自为 （自动统计）	代码化 计算
产:	DXsjk	>	GXsjk	人教机为 （一劳永逸）	标准化 复制
学:	DXsjk	<	GXsjk	机教人为 （思维球茎）	载体化 共享
研:	DXsjk	?	GXsjk	人机共为 （默契通信）	法制化 交换
用:	DXsjk	<	GXsjk	机代人为 （一通百通）	自动化 复用

图 3

图4是本发明的知识工程示意图。

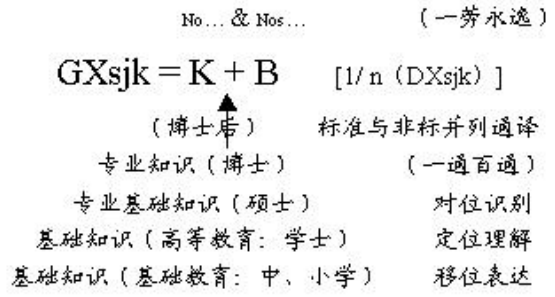


图 4

图5是本发明的产权监管示意图。

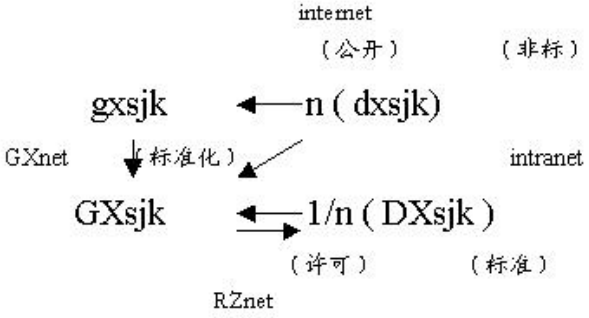


图 5

图6是本发明的默契通信示意图。

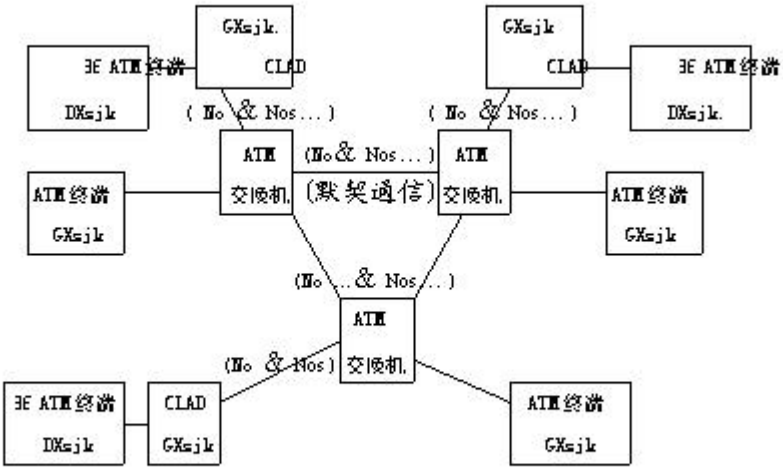


图 6

本发明的有益效果或好处是：

建立基准参照系及应对参照系，解决因知识信息错位而引起的冗杂文本、非标形式、垃圾信息、知识爆炸和怪圈悖论等难题。

打个比方来说，上述这些难题就好比是一团乱麻。采用现有理论的范式和现有技术的办法就是：要么是快刀斩乱麻，这是从整体介入的通常作法；要么是慢慢地去理，这是从局部介入的通常作法；要么是：先把这团乱麻分成一堆一堆的小团，再分别从各个小团的内、外两个方面逐步深入地去理或斩（注：最后能否解决这团乱麻带来的所有问题，就要另当别论了！一方面，要看这团乱麻本身乱的性质及程度，另一方面，则要看你是否真正想要解决问题或打算把问题解决到什么程度），这是先整体、后局部，乃至不断往复逐步深入的方法。

本发明采用的方法则是：好比是先给这团麻绳一根一根地染上色，有的甚至再编上号（即：从全域基因文本元素的统一编号入手），然后，再针对具体情况（如：已知域基因文本组合）并考虑实际需要（如：目标域基因文本组合），既有根据又有针对性地，选择不同的颜色及编号（如：确定具体的参照系乃至参照系的组合），或分、或斩、或理（悉听尊便！）。

这实质上指出了解决知识信息的计量及测度难题的最佳途径，并且，从方法及产品两方面给出了简明的示范。例如：基准参照系的应用及其效果，充分体现了人工智能主体以高速运算和海量存储支持的演绎及完全归纳优势，表现为对数据的狭义处理和广义处理的发散或收敛的精密性或专一性；应对参照系的应用及其效果，则充分体现了人类智能主体以跨时空的联想、想象、灵感、直觉支持的类比及非完全归纳优势，表现为对知识或信息的狭义处理和广义处理的发散与收敛的易变性或多元性（实质上也就是粗放性或通用性）。有必要通过协同运行，使它们优势互补；从载体方面看，就是要克服物理载体的机械性和生物载体的易变性，即实现人机之间的取长补短。

由此可见，本发明的重要性、必要性和可行性，不仅易于理解而且容易证实。

如果能够逐方面、逐阶段、逐层次、逐系统地实施本发明，那么知识信息的计量及测度这一国际难题的解决必将成为现实。那时，人类利用知识信息的整体水平和综合效能必将显著提高。进而，必将促使人工智能与人类智能的优势互补，形成效率更高的协同智能，促成形式信息革命向语义信息革命的时代飞跃。

从性能方面看，本发明，既可在信源端一劳永逸地解决非标形式的识别难题，又可通过信道实现

默契通信以及借助交换终端实现一通百通的形式变换（包括多种语言乃至多元基因文本形式的通译）排除冗杂文本的干扰，还可在信宿端利用一元基因文本直接走短程捷径（防止或避免垃圾信息、知识爆炸和怪圈悖论），进一步，还可利用本发明对广义真实文本中涉及的知识信息进行自动测评，包括数量计算和价值评估。

总之，本发明通过人工智能与人类智能、电脑与人脑、电信网与神经网的优势互补，不仅能够显著地提高协同智能主体的效率及效能，而且能够显著地改进现有技术这种在总体上仍是高消耗、低效率的知识信息处理方式。因此，本发明不仅是工业化向信息化转换的利器，而且是形式信息革命阶段向语义信息革命阶段转变的利器，其应用涉及算、产、学、研、用等各个方面，适用于一切需要加速信息化进程以及需要进一步提高信息化工程的质量的国家或地区的法人及自然人。

实施例1

全域数码化网络，是由码、卡、表、库、网、端构成的融智网络（RZ-net），其中，码与其它具体的式之间具有对应转换关系，整个系统表现为纯数字形式和纯载体形式的交集，其特征在于：以码代式进行知识信息数据处理；纯数字形式，采用（a+bi&…）的形式语言编制源程序，以（a+bi&…）与（0&1）的交集作为编译程序；纯载体形式，采用卡、表、库、网、端的具体形式承载码。

纯数字形式，由基准参照系和应对参照系构成。其中，包含纯数字系统的全域基本元素通式与元素组合特式的演绎法则，即：全域基本元素0、1、2、3、4、5、6、7、8、9构成基准参照系，其它所有的数字都可由前述阿拉伯数字的组合构成，包括根据已知域运算符号及公式等数学法则及原理由基本元素衍生或派生各种各样的组合形式，涉及相应的时间程序和空间结构以及它们的整合架构或框架，各个具体的应用系统选择相应的目标域基因文本组合形式体系构成应对参照系。只要识别上述通式或基准参照系，同时，又选择适当的特式或应对参照系，就能使电脑与人脑的优势互补，而且能使一劳永逸地复用已建立的基准参照系及应对参照系。

实施例2

全域数码化出版物，是由码 + 式、卡、表、库、网、端构成的融智出版物（RZ-print for client/server），由于码与其它具体的式的对应转换或联系，整个软件系统表现为纯数字系统的子集与其相应的具体的式的子集的交集或并集，如：以（bi&…）表示的图、文（注：狭义的文）、数、表、音、像等各种具体的基因文本形式，其特征在于：码式并行的知识数据处理方式，即从广义真实文本中提取的多元基因文本元素及其组合构成的基准参照系及应对参照系，与纯数字系统中相应数元的基准参照系及应对参照系一一对应，其中，码式并行还表示每一个具体的基因信息元素及组合都是被相应数元的码式并行地为卡所承载的。

因为义、文（注：广义的文）、物、意，分别与（a＋bi＋cj＋dk）一一对应，所以图、文（注：狭义的文）、数、表、音、像，分别与（bi&…）一一对应。其中，广义的文，包括图、文、数、表、音、像；狭义的文，包括各种各样的语言文字或符号体系。

由此可见，实施例1与实施例2是“编号排位”与“对号入座”的关系。

全域数码定位方法，就是给基因信息及基因文本“编号排位”。其步骤是：（1）“编号”：在（a＋bi&…）代码体系与（0 & 1）数字体系之间建立一一对应关系，从而，构成全域数码体系，即本发明的系统软件；（2）“排位”： 在码与卡之间建立一一对应关系，进而，构成卡、表、库、网、端一体化的全域定位体系，即本发明的硬件形式，其中，卡为码的物化展示单位，表为码的物化操作界面，库、网、端是码的集中、分布、整合的物化结构形式。如：实施例1的纯形式系统。

纯文本系统及纯数码系统，就是在纯形式系统中，对已知域及目标域真实文本涉及的基因信息及基因文本进行“对号入座”。 也就是在融智网络中使其它基因文本元素及组合“各即各位”。其步骤是：（1）“对号”：在全域数码体系中给其它多元基因文本元素及组合进行代码定位；（2）“入座”：在全域定位体系中给其它多元基因文本元素及组合进行载体定位；（3）从真实文本中提取已知域基因文本，构成集大成共享基因文本（包括：数码文本）数据系统；剪接或重组目标域基因文本，构成集小成独享基因文本（包括：数码文本）数据系统。如：实施例2及实施例3。

动、静态一体化的融智出版物，主要有两种产品类形：（甲）是常例，即已知域集大成共享广义文本，通常以融智网络交换端服务器共享数据库（server：GXsjk）的形式发挥集中式全域数码化多媒体图书馆或百科全书的作用；（乙）是特例，即目标域集小成独享广义文本，往往以用户端计算机或其它终端机独享数据库（client：DXsjk）的形式发挥分布式全域数码化多媒体百科全书（如：摘要、选集、分册、简写本等个性化图书）的作用。

一方面（甲）是由（乙）聚汇并合并同类项而成，另一方面，（乙）是对（甲）有选择的复制或复用；（甲）和（乙）都可依托融智网络的支持，对真实文本协同实施域位解析，从而，均可自建相应的全域数码化广义文本数据系统；不仅（甲）与（乙）之间，可互为信源和信宿，而且，（甲）的子集之间或（乙）的子集之间，也都可以互为信源和信宿。

通过在纯数字系统相应数元的基本元素通式及元素组合特式与其它相应的多元基因文本元素通式及组合特式之间建立一一对应关系，即可实现任意一个基因信息元素及组合的同义的多元基因文本元素及组合相互之间的一通百通或并列通译。只要识别通式或基准参照系，同时，又能选择适当的特式或应对参照系，就能把电脑的演绎能力与头脑的直觉能力相结合使本发明产品在处理知识信息时发散与收敛得更到位，从而，可以准确无误地判定每一个知识点及表达式在整个知识及基因文本体系中的地位、作用和意义。

如果说知识系统工程的任务是建立其它相应的多元基因文本元素通式及组合特式，那么文化基因工程的任务就是在纯数字系统相应数元的基本元素通式及元素组合特式与其它相应的多元基因文本元素通式及组合特式之间建立一一对应关系。实施例2关于知识系统工程的任务及对策或办法举例说明如下：

1、以基本几何图形构成纯图形文本的基准参照系，根据已知域几何图形组合及其变换法则，派生出更多的更复杂的组合形式，各个具体的应用系统选择相应的目标域基因文本组合形式体系构成纯图形文本的应对参照系。

应用示例：建立广告图案(特别是其中的商标类)、工业制图、电子地图、建筑制图、模型图、速写、素描、油画、国画等具体的基因文本共享数据库。

2、以基本字母或基本笔画构成纯文字文本的基准参照系，根据已知域字母或笔画组合（如：词素、词、词组、句、段、篇、章）及其原理或法则以及相应的标点符号及使用法则，派生出更多的更复杂的组合形式，各个具体的应用系统选择相应的目标域基因文本组合形式体系构成纯文字文本的应对参照系。

应用示例：通过对用现有技术制作的字典、词典和各种各样的语料库进行改造使之全域数码化，建立词素、词、词组、句、段、篇、章等各种语言文字的基因文本共享数据库。除此之外，特别需要对各个学科、各个产业、各个领域的知识信息文献（包括教科书、专业著及论文、普及读物等）进行全域数码化改造，建立版权类、Know-why类、专利权类、Know-how类的基因文本共享数据库。

3、以基本数字（如：阿拉伯数字0、1、2、3、4、5、6、7、8、9）构成纯数字文本的基准参照系，根据已知域数字组合（如：不同数系的数以及数学形式）及其原理或法则以及相应的数学符号及使用法则，派生出更多的更复杂的组合形式，各个具体的应用系统选择相应的目标域基因文本组合形式体系构成纯数字文本的应对参照系。

应用示例：建立从小学到博士后不同层次的几何类、代数类、分析类的基因文本共享数据库。

4、以基本表格或图表构成纯表格或图表文本的基准参照系，根据已知域表格或图表组合及其原理或变换法则，派生出更多的更复杂的组合形式，各个具体的应用系统选择相应的目标域基因文本组合形式体系构成纯表格或图表文本的应对参照系。应用示例：建立各种程序、结构以及框架的简表及详表共享数据库。

5、以基本音素或音阶构成纯声音文本（语音或音乐）的基准参照系，根据已知域音素或音阶组合（如：音节及言语片段或旋律及乐曲）及其原理或法则，派生出更多的更复杂的组合形式，各个具体的应用系统选择相应的目标域基因文本组合形式体系构成纯声音文本（语音或音乐）的应对参照系。

应用示例：建立国际音标、音节、不同语言的词语以及言语片段等版权类基因文本共享数据库；建立不同音调的音阶以及各种各样的旋律及乐曲等版权类基因文本共享数据库。

6、以基本透视图及三视图和基本色彩及色调构成纯视像文本的基准参照系，根据已知域视图组合和色调变换（如：各种自然物及人工物的立体或活体影像）及其原理或变换法则，派生出更多的更复杂的组合形式，各个具体的应用系统选择相应的目标域基因文本组合形式体系构成纯视像文本的应对参照系。

应用示例：建立广告、工业制图、电子地图、建筑制图、模型图、教学演示、经济模型、产品模型、虚拟模拟、电影电视节目等版权类基因文本共享数据库。

7、以基本粒子构成纯立体文本（物质）的基准参照系，根据已知域粒子组合（如：原子、分子以及各种具体的物质）及其原理或法则，派生出更多的更复杂的组合形式，各个具体的应用系统选择相应的目标域基因文本组合形式体系构成纯立体文本的应对参照系。

应用示例：建立全域数码化博物馆、文物馆或历史馆、太空馆等共享系统。

8、以碱基构成纯活体文本（生物）的基准参照系，根据已知域碱基组合（如：密码子、氨基酸以及各种具体的生物）及其原理或法则，派生出更多的更复杂的组合形式，各个具体的应用系统选择相应的目标域基因文本组合形式体系构成纯活体文本的应对参照系。

应用示例：建立全域数码化生物馆、标本馆等共享系统。

实施例3

全域数码化终端，是由式 & 码、卡、表、库、网、端构成的融智终端（RZ-client/server），整个系统表现为多元基因文本的一元数码文本形式，其特征在于：以码代式进行信息数据处理，在融智网络支持下，以动态形式的融智出版物的简式或省略形式进行远程协同操作及知识数据处理。

由此可见，实施例3是以实施例1和实施例2为依托进行知识产权监管和默契通信的全域数码化终端，主要以适合与其它终端进行直接通信的形式或方式发挥整合式全域数码化多媒体情报站的作用。如：随时“对号入座”地复用并复制新的或发现并公开新的基因文本元素及组合；在端与端之间互通有无或默契通信。

由于依托融智网络及融智出版物，融智终端可随时复制或复用任何领域的知识信息数据或基因文本，因此，不仅交换服务终端可以根据一元数码文本再现并提供相应的多元基因文本给用户，而且用户终端也只须存储有限的个性化信息数据（而无须重复存储冗余的知识数据，特别是通常无须重复存储冗余的多元基因文本，需要时只须以简式或省略形式）就能够进行海量的知识信息数据处理。融智终端不仅能自动对付冗余文本或垃圾信息干扰，而且能在基因文本的海洋中发现短程捷径。

以下结合附图和实施例，对本发明作进一步说明：

在图1中，（a + bi &…）表示多元数系的一般形式，a0+{ a1，a2，…}表示实数或实部的一般形式a0+{ a1，a2，…}，（bi &…）表示虚部的一般形式。多元数（a + bi &…），由实部（a）和虚部（bi &…）两部分构成。实部或实数（a），指代本真信息，其中，每一个具体的数字（a）都表示基因信息元素(涉及已知域及未知域)在实数集合中的一个特定序位；虚部（bi &…）对应于多元基因文本，每一虚数元（bi）都标识一类基因文本，如：标识图、文、数、表、音、像（包括立体与活体的文本形式），进一步包括标识多语种、多物种或多学科（注：每一个学科都有一套相应的形式语言）。

全域数码定位系统，是由基准参照系及应对参照系构成的全域数码化网络。其中，基准参照系，由全域基因文本元素0、1、2、3、4、5、6、7、8、9构成；应对参照系，通过选择适当的目标域基因文本组合构成，而各种各样的基因文本组合则是根据体现已知域相应的法则的其它数学符号和基因文本元素共同衍生出来的。作为确定具体基因文本的定位工具的参照系，其中，每一个数码都有一个特定的序位，每一数段都对应于一个基因信息元素（a）和多类基因文本元素（bi &…）。

本发明对语义信息及真实文本的量化处理，就是借助文化基因通式的简式即多元基因文本通式（a + bi &…）对特式或特例的解析而实现的。

在图1中，已知域（m，n）和未知域（x，y），涉及指代义、文（注：物可以被视为立体与活体文

本，也可以用动态的音、像文本虚拟表示）的全域数码；目标域（a，b），涉及意——智能主体选择基因信息（即本真信息）及基因文本（即形式信息）的过程及状态。（a，b）尽管通常只是（m，n）的子集，但也可是（m，n）的并集，还可涉及（m，n）与（x，y）的交集。（m，n）与集大成共享基因文本数据库GXsjk属于同一范畴；（a，b）通常与集小成独享基因文本数据库DXsjk属于同一范畴。基因文本元素集合是完全归纳集，基因文本结构集合则多数是相对完全归纳集（因为它也有一部分是完全归纳集）。

全域数码集合具有完全归纳（或完全演绎、或完全类比）的特性。

设：a = b + c + d +… ，则：a 〉b & c & d &… ，即：本真信息元素集合是唯一的全集，属于完全归纳范畴，广义文本的多元基因文本各子集即基因文本元素集合也属于全归纳范畴。又：（a + bi）&（a + cj）&（a + dk）&… ，或：（a + bi）=（a + bi & cj & dk &…），所以，（a + bi）是（a + bi + cj + dk + …）或（a + bi &…）的复部，即：复数域暨复平面是智能主体出入多元数系暨多维空间（包括四维时空）对文化基因及其代码进行选域定位的转换界面。

本发明通过一系列量化转换界面出入多元数系暨多维空间对基因信息及基因文本的元素及组合进行选域定位。选域是选择虚数元，从而识别基因文本类别或通式；定位是在通式中确定具体代码的位置、位移及轨迹，体现为一系列代码组或特式。

图1表示全域数码文本形式的文化基因通式（a + bi & cj & dk）的最简表达式（a + bi &…），这一数学模型集几何类、代数类、分析类于一体，涉及全域数码的异义排列、同义并列、经纬阵列、多维选列，在图2中表现为物理模型即量化转换界面。文化基因通式中包含许多具体的多元基因文本类型（bi &…）。

同一基因信息的多元基因文本（包括标准与非标形式），可以通过同义并列实现通译或对应转换。无论是借助一元数码文本通式及其量化转换界面，以最简单且最经济的方式，直接处理本真信息（即：走思维捷径），还是选择具体的多元基因文本通式及其量化转换界面，以具体的用户最熟悉的形式，间接处理衍生形式或广义文本（即：寻找联想短程线），都必须依托统一的文化基因通式，因为它的全域数码包含了所有的多元基因文本通式（包括已知域及未知域）。

具体智能主体的DXsjk或（a，b）的聚会，可简化通约为GXsjk或（m，n）。（m，n）逼近（x，y）的过程是通过用户对相应的（a，b）具体操作实现的。借助共享网络（GXnet）和相应终端的协同运行，通过具体的计算、复制、复用、交换（或交流）、共享，任何一个DXsjk与GXsjk之间，都能够在全域数码化的条件下进行比较甚至互换。选域定位、协同操作和去冗存要，就是建立在这种可比性的基础之上的。正是基于此，对语义信息及真实文本的定性分析、定量分析和结构分析，才能通过文化基因通式及其量化转换界面对每一特式或特例的定位分析而实现。

在图2中，活页卡（K）和活动表（B）构成文化基因通式的最简物理模型。自动统计基因信息元素与自动计量基因信息组合，是通过基因文本元素及组合（如：结构、程序、框架等）的复用次数（注：统计全域数码和计量全域数码组）或复制件数（注：计算活页卡的数量）的自动统计及自动计量而实现的。

码、卡、表，是计量并操作基因文本元素及组合的基本工具。人（个、群、类）关于各物种、各语种、各学科的知识（意）与宇宙万象（物、文）及其机理或其运行法则（义）是否一致？只有在文化基因通式中通过对基因信息及基因文本的协同量化处理，才能做到既见树木又见森林。码、卡、表、库、网、端的有机统一，使智能主体可以对语义信息及真实文本进行高效率、低消耗的协同量化处理。

本发明对广义真实文本中的基因文本元素及组合的选域定位、协同操作和去冗存要，就是通过一系列记录并承载基因信息元素及组合的量化代码及活页卡的处理而实现的。

图1和图2是通过实施例1从纯形式的角度说明本发明。

由码、卡、表、库、网、端构成全域数码定位系统是完整的产品形式，包括：

（1）全域数码系统软件（其中的量化转换界面是协同操作系统）及硬件，由实数代码及虚数标识与二进制数字一一对应而构成的全域数码化网络；

（2）基因文本元素子系统，是存储与全域数码体系各子集相对应的图、文、数、表、音、象等基因文本元素的原型或样品的代码数据系统。它是以码、卡、表的形式表达的基因文本元素及组合的共享全域数码数据库（与实施例2所述的多元基因文本数据库相对应）及硬件，即全域数码化出版物，其内容是编译程序（其各个子集则分别与相应的多元基因文本元素构成的各个全域一一对应）；

（3）用户应用子系统，是用户选择的已知域基因文本系统（如：集大成共享基因文本数据库和共性化共享基因文本操作系统及载体）或目标域全域数码系统（如：集小成独享全域数码数据库和个性化独享全域数码操作系统及载体）的全域数码化终端。

在图3中，根据集小成独享数码文本数据库（DXsjk）与集大成共享基因文本数据库（GXsjk）之间的关系（即：它们所指代和承载的基因信息之间的等量关系，通过计算相同的全域数码及全域数码组的数量而判定），从算、产、学、研、用几个方面，说明本发明的以下基本技术性能及功效。

1、计算全域数码组即基因文本组合，当DXsjk等于GXsjk时，机器自为，即以自动统计的计量方式，进行代码化文化基因的演绎或完全归纳或完全类比推理。

2、复制全域数码组即基因文本组合，当DXsjk大于GXsjk时，人教机为，即以一劳永逸的生产方式，由用户指导机器学习——基因文本标准化，即对位识别。

3、共享全域数码组即基因文本组合，当DXsjk小于GXsjk时，机教人为，即以短程捷径的学习方式，由机器启发用户顿悟——基因信息载体化，即定位理解。

4、交换全域数码组即基因文本组合，当DXsjk不等于GXsjk时，人机共为，即以默契通信的研究方式，调适人机各方直觉——主体行为法制化，即约定交流。

5、复用全域数码组即基因文本组合，当DXsjk小于GXsjk时，机代人为，即以一通百通的应用方式，有关子系统协同联想——选域定位自动化，即移位表达。

图4表示现有科教知识体系支持的知识工程与本发明的文化基因全域数码文本通式支持的文化基因工程之间的关系。

在图4中，全域数码（No.）及全域数码组（Nos.）分别指代多元基因文本元素及组合。全域数码化的活页卡（K）专门承载基因文本元素及组合。箭头表示各个方面、各个阶段、各个层次和各个系统的知识信息以全域数码化和活页卡片化的形式进入集大成共享基因文本数据库或集小成独享数码文本数据库。协同智能主体通过活动表（B）展示量化处理语义信息及真实文本中的基因信息及基因文本的过程，如：对本真信息一劳永逸的复用、对基因文本一通百通的复制、对标准与非标形式的基因文本的并列通译。1/n是代数符号。“=”表示GXsjk与n（DXsjk）在网络的支持下基因信息总量的动态守恒。

由基础知识、专业基础知识、专业知识等构成的层次阶梯，表示以码、卡、表、库、网、端的形式对现有知识体系进行全域数码化改造的内容。

图4通过实施例2从多元基因文本的角度说明本发明。

图5和图6通过实施例3分别从网络（整体）与终端（局部）两方面说明本发明，涉及网络知识产权监管（图5）和终端之间的默契通信（图6）。

图5表示由已知域及目标域基因文本支持的网络知识产权监管体系。在图5中，n 和1/n是代数符号。共享数据库（gxsjk）、独享数据库（dxsjk）、国际互联网（internet）和局域网或内联网（intranet）是现有技术。它们均采用真实文本与二进制数字一一对应的数据转换模式或处理方式。由二进制数字和逻辑电路构成的共享通信网（GXnet）是连接现有技术与本发明技术的纽带。本发明采用全域数码及全域数码组与逻辑电路一一对应而构成的协同智能通信网或融智网络（RZ-net），是以集大成共享基因文本数据库（GXsjk）和集小成独享基因文本数据库（DXsjk）为支持的知识信息数据处理系统。等号表示在连网的条件下GXsjk与DXsjk可以是并集。本发明对网络知识产权的监管，是通过对全域数码及全域数码组的监管而实现的。对已知域、未知域、目标域的基因信息及基因文本的知识产权监管都是以公开或许可的方式体现的。其步骤是：首先，从真实文本中提取基因文本；然后，再将基因文本全域数码化，以便对复用或复制的数量进行统计或计算。

这就把网络中繁杂的语义信息及真实文本的知识产权监管简化为对全域数码及全域数码组的自动

统计和监管。

图6表示由已知域及目标域基因文本支持的各终端之间的默契通信。

在图6中，信元装拆设备（CLAD）和异步转移模式（ATM）是现有技术。它只传输与真实文本一一对应的信元及信息段（即：二进制数字），而与文化基因全域数码及全域数码组无关。本发明技术则传输与全域数码组一一对应的二进制数字，通常不必直接传输全域数码文本以外的其它多元基因文本，更不必直接传输真实文本。信道只传输二进制数码，终端才存、处全域数码组（Nos.）以及相应的多元基因文本，并根据用户需要再现真实文本或重构目标文本。用户可以根据实际需要，至少在以下三个层次实现全域数码组以及基因文本组合与真实文本之间的对应转换，即：在交换机之间；在交换机与终端机之间，包括在CLAD与非ATM终端机之间；在终端机之间。

本发明的又一种表述如下：

本发明是处理语义信息及真实文本的文化基因工程方法及产品，具体包括：

一、文化基因工程方法，其步骤是：

1、在复数代码体系与二进制数字体系之间建立一一对应关系，从而，构成对指代文化基因的全域数码及全域数码组进行全域定位的数学模型。它涉及已知域及未知域，相当于一个拥有无穷多个座位及编号的超级剧场的“号码体系”；

2、以卡为单元载体，在码与卡之间建立一一对应关系，用表对之进行操作，以库、网、端的形式，构成对承载基因文本的元素卡及结构卡进行全程扫描的物理模型。涉及目标域，它相当于上述超级剧场的“座位体系”；

3、由码、卡、表、库、网、端，构成全域数码定位系统，如：实施例1，它相当于该超级剧场的整个号码座位体系；从真实文本中“对号入座”地自动提取相应的基因文本，从而，构成自由剪接或任意重组已知域及目标域基因文本的柔性加工用户模型，如：实施例2和3，它们相当于该超级剧场的以下“入场情形”，即：“实际入场”（有“入场者”，即已知域其它多元基因文本）和“虚拟入场”（无“入场者”，只有目标域其它多元基因文本的代码）两种情形]。

二、文化基因工程产品的三个实施例：（1）由“号码体系”和“座位体系”构成的纯形式系统产品，它相当于该超级剧场的整个号码座位体系；（2）由“实际入场”的已知域及目标域其它多元基因文本构成的纯文本系统产品；（3）由“虚拟入场”的已知域及目标域全域数码文本构成的纯数码系统产品。

域、位、点、式、码、卡、表、库、网、端的含义及特征：

域，即范围，在此是指：基因信息及基因文本元素构成的全域，即文化基因通式及基因文本通式涉及的范围，包括由派生的各种各样的基因文本组合构成的已知域、未知域和目标域（注：按智能主体对基因信息及基因文本的认知情况而划分）。

位，即位置，在此是指：基因文本元素及组合在基准参照系及应对参照系中的位置或序位。

点，即知识点，在此是指：基因信息元素及组合。

式，即表达式，在此是指：多元基因文本元素及组合。

码，即数码，在此是指：全域数码，包括多元数及其二进制数字形式，其特征是：由（a+bi&…）及其（0&1）形式体系构成全域数码体系，并用它作为协同智能主体的标准语言或符号体系，同时，以复数系及复平面或曲面（a+bi&…）的形式作为出入多元数系及多维时空的量化全域切换界面（使基因文本元素可控的基准参照系成为操作界面），通过时空变换产生无数的量化子域切换界面（使基因文本结构或程序以及框架可控的应对参照系成为操作界面）。其中，用实数指代异义排列的基因元素，用虚数标识同义并列的文本形式，用多元数表示多维阵列的多种基因文本。把已知域的其它多元基因文本简化为相应的全域数码文本，以便于人机协同处理，必要时再还原为相应的真实文本（通过基因文本通式给其它各种形式体系及元素选域定位——自动化的识别、理解、表达或通译）。

卡，即活页卡，在此是指：单元化操作页面，包括电子和非电子形式，其特征是：元素卡用于承载多元基因文本元素，附有异义单列的一个实部代码和同义并列的某个虚部标识；程序卡是元素卡的时间序列形式；结构卡是元素卡的空间排列形式；架构卡是程序卡和结构卡的综合形式。

表，即活动表，在此是指：多元化操作界面，即各个终端协同处理知识信息和协同操作载能载体的传感传播界面，其固化形式可是通用和专用芯片或光磁盘等，如：全域数码化的视窗操作界面及其各种应用软件，其特征是：它是由全域数码化的元素卡、程序卡、结构卡、架构卡构成的多码多卡多维阵列（包括经纬阵列或多维选列），附有异义排列的多个实部代码和同义并列的多个虚部标识；用于逐方面、逐阶段、逐层次、逐系统地展示文化基因，并对之进行各种各样的排列组合。

库，即数据库，在此是指：是由图、文、数、表、音、像等多元基因文本元素及组合构成的全域数码化出版物，其特征是：以基因文本替代真实文本作为存储对象，而且，所有的基因文本元素及组合都全域数码化。

网，即通信网，在此是指：是由全域数码体系和全域定位体系构成的融智网络，其特征是：以全域数码替代真实文本乃至其它多元基因文本作为传输对象。

端，即终端，在此是指：是既能以用户最熟悉的某种文本形式进行人机对话，又能在各个终端之间进行端到端默契通信的终端，包括：交换机、服务器、用户计算机、数字电话机、数字电视机、数字家电、传感器、终端芯片、IC卡乃至终端标识（如条形码和非电子化的活页卡）等等各种具体的终端形式，其特征是：以全域数码替代真实文本乃至其它多元基因文本作为处理对象。

上述十个方面的相互关系：

如果把本发明的产品形式比喻为一个超级剧场，那么，域，就是该剧场所有的座位及其编号可能涉及的任何一个范围，涉及各个实在或虚拟的分剧场；位，就是该剧场的各个被实在或虚拟地设置或占用的序位；点，就是该剧场的座位及编号被实际或虚拟地占用的具体分布情形；式，就是该剧场的座位及编号被占用的情形的称谓或叫法，涉及各种称谓体系；码，就是该剧场的座位的编号，它是上述所有的称谓体系中的一种既最全面又最简捷的标准称谓体系或全域数码体系；卡，就是该剧场的最简捷的标准座位形式；表，就是该剧场的座位及编号指南或视频向导；库，是该剧场的座位编号的集中形态；网，是该剧场的座位编号的分布形态；端，是记录该剧场占用情形的装置，包括出入口。

本发明方案涉及的部分创新概念的有关词语的基本含义说明如下：

1、文化基因，是从与物化基因（如：生物基因）对应的观点提出的概念。它包括基因信息及基因文本。文化基因的提取，是以活页卡的载体形式和全域数码的文本形式进行的，实质上就是从广义真实文本中提取其它多元基因文本并对之进行文面表达。文化基因的剪接或重组，是以活动表的界面形式进行的，实质上就是对基因文本组合中各个元素的序位进行剪辑或重构。文化基因的提取、剪接或重组的过程，也就是各个智能主体对基因信息及基因文本进行选域定位、协同操作和去冗存要的过程，即在用全域数码文本表达的文化基因通式中，对具体的基因信息及基因文本的位置、位移及轨迹进行全域数码定位。文化基因的提取、剪接或重组的结果或成果，或集中、或分布、或整合，以便于为用户复用或复制（含交流与共享）。

2、基因文本元素代码，是指代或替代其它多元基因文本元素的代码（a + bi&…），如：图、文、数、表、音、像等各类（bi&…）基因文本元素代码。

3、已知域基因文本元素，是智能主体（包括：人类智能、人工智能、协同智能等主体形式）已知范围内的基因文本元素，如：图（基本几何图形）、文（拼音字母，汉字的基本笔画）、数（如：阿拉伯数字符号）、表（基本表式）、音（音素，音阶）、像（基本的三视图，三原色）。

4、异义排列、同义并列、经纬阵列、多维选列，是指在文化基因通式中，基因信息的异义排列（a）、基因文本的同义并列（bi &…）、文化基因的经纬阵列（a + bi）或多维选列（a + bi &…）。

5、基因文本组合代码，是相应的基因文本元素代码的组合。

6、对位识别、定位理解、移位表达、域位解析，是指在文化基因通式（a + bi &…）中借助多元转换界面对其它具体的多元基因文本进行“对号入座”[区别于“不对号入座”（即：入场者座错位置）或“对号不入座”（即：虚拟入场）的情形]的过程。

7、全域数码，是（a + bi &…）代码集合与（0 & 1）数字集合的交集的元素。采用复数形式的目的是对基因信息及基因文本进行全域定位，采用二进制数字或数据形式的目的是便于人工智能主体

对之进行（广义的）处理。

8、去冗存要、去冗处要、去冗传要、去冗馈要，是由广义真实文本到基因文本（包括全域数码），以及由未知域到已知域再到目标域的收敛过程，涉及对知识信息数据的广义处理（即：输入、存储、处理、输出、传输、反馈）和狭义处理[即：分解与合成（含：重构和重组）]。

9、协同智能，是人工智能与人类智能的协同运行方式。协同智能主体是智能主体的一种进化形态。

10、基因文本元素，是构成基因文本组合的基本元素，如：图、文、数（含全域数码）、表、音、像等各个形式体系的基本元素。

11、全域定位，是指：基因文本元素在其所属的全域中都有特定的序位，所有因各个基因文本元素的时空变换而构成的基因文本组合（涉及已知域、目标域、未知域），在基因文本通式中都有各自相应的序位记录。

12、协同操作，是以量化转换界面（表），把指代已知域图、文、数、表、音、象，以及各语种、各物种和各学科的基因文本，用码和卡进行分解与组合，以库、网、端的形式，进行集中、分布、整合的处理，以便人机协同探索未知域。

13、基因文本组合，是构成广义真实文本的各种基因文本程序、结构、框架，它本身则由基因文本元素所构成，如：图、文、数、表、音、像等各个形式体系的不同层次的程序或结构及框架。

14、横断扫描，是利用全域转换界面，对广义真实文本中包含的基因文本，进行逐方面、逐阶段、逐层次、逐系统的全域扫描或有针对性的搜索。

15、分离与组合，是系统或用户利用全域转换界面，借助码和卡对广义真实文本中的基因文本，进行提取、剪接或重组的基本操作。

16、去冗存要，是指集小成独享基因文本或用户终端（可以是交换机、服务器、计算机乃至其它各种具体的个性化终端等等形式）从广义真实文本中提取基因文本并全域数码化的过程。

17、全域数码文本，是全域数码及全域数码组这一特殊的基因文本形式。

18、柔性加工，在此是指系统或用户借助基因文本对真实文本的随意加工。

19、文化基因通式，是由义、文、物、意四元素构成的融智概念体系的基因形式，其数学形式涉及：复数、四元数（特例）乃至其它多元数的相应形式。

20、人机优势互补，是指：一、人机两方面在知识信息数据处理领域的优势互补，如：推理、学习、顿悟、直觉、联想等人类智能的优势与计算、复制、共享、交换、复用等人工智能的优势的协同互补；二、人机两方面在载体载能形式变换领域的优势互补，如：生物载体的易变性等人类智能主体的特点与物理载体的机械性等人工智能主体的特点之间的协同互补。

21、自动测度（包括自动统计和自动计量），是指本发明对基因文本在全域中的序位记录（涉及复制、复用、共享、交换的数量）进行的自动计算。如：在整个网络中对某些全域数码组或基因文本组合的实际使用和交换数量的自动测量。

22、一劳永逸，是指只要复制一次就能多次复用。

23、短程捷径，是指一劳永逸地共享已知域多元基因文本。

24、默契通信，是指主体之间心有灵犀一点通的默契交流在全域数码化网络以及出版物或终端中的体现，即：实部编号确定本真信息，虚部标号确定多元基因文本的类型，根据全域数码即可再现已知域多元基因文本及目标域真实文本。

25、一通百通，是指全域数码化基因文本一通百通，因为表示同一基因信息的多元基因文本实部编号一致，所以基因文本只要全域数码化就可以随时随地转换成为已知域多元基因文本的其它任何一种形式。

26、本发明的应用涉及电信网（以tel.表示）、广电网（以TV表示）、计算机网（以computer表示）、出版物网（以CPU，IC，CD-room or paper表示）。

注释　　系统科学之窗 论文专区（2000-12至2005）

该发明已于2000年11月29日在国家知识产权局的中国专利《发明公报》第16卷48期提前公开，申

请号是001093800，公开号是CN1274895A，全文20088字，六幅图解。发明人兼著作权所有人：邹晓辉

一种知识信息数据处理方法及产品

【申请号】 【公开号】	CN00109380.0 CN1274895	【申请日】 【公开日】	2000-05-31 2000-11-29
【申请人】	邹晓辉	【地址】	519125 广东省珠海井岸桥东恒美花园 15－2 栋 201 房
【发明人】	邹晓辉		
【国省代码】	44		
【摘要】	本发明属于语义信息及真实文本处理技术。本发明的目的是：提供一种知识信息数据处理方法及产品。本发明方法是文化基因工程方法，特征是：从广义真实文本中提取、剪接或重组文化基因；本发明产品是全域数码定位系统,由码、卡、表、库、网、端构成,特征是：全域数码化的网络、出版物或终端。本发明的有益效果是：建立基准参照系及应对参照系,解决因知识信息错位而引起的冗杂文本、非标形式、垃圾信息、知识爆炸和怪圈悖论等难题。		
【页数】	22		
【主分类号】	G06F17/22		
【专利分类号】	G06F17/22;G06F15/163		

结构“树”——为帮助理解而(1999-2001)做的内容详解

- 一种知识信息数据处理方法及产品
  - 联系电话：0756-5505041
    - email: qhkjy@yahoo.com.cn
      - 作者及发明人：邹晓辉
        - 推广成果期间，现任：清华科技园（珠海）融智文化基因工程研究所（筹）所长
        - 公开该成果时，曾任：吉林大学国际交流学院珠海分院 科教处主任

- 发明概述
  - 属于：一种语义信息及真实文本的数字处理技术，进一步是一种知识信息数据处理方法及产品。
  - 涉及：人工智能、计算机和通信技术的交叉综合领域。
  - 问题：人们不能从根本上解决冗杂文本、非标形式、垃圾信息、知识爆炸和怪圈悖论等难题。
  - 原因：人类关于语义信息的定性分析、定量分析和结构分析长期未能获得实质性的重大突破。
  - 需要：在人工智能与人类智能、电脑与人脑、电信网与神经网之间实现协同效能或优势互补。
  - 现状：现有技术能对受限范围的知识信息及真实文本进行局部的量化处理，在非受限范围一筹莫展。
- 当前的语义学理论和语义信息处理理论存在的问题
  - 对语义信息的本质认识或阐述不清
  - 把形式信息与语义信息混为一谈
  - 把衍生形式与本真信息混为一谈
  - 不知道语义信息的本质究竟是什么
  - 语义信息理论体系是一个大杂烩，根基不牢固
  - 未能明确地区分本真信息、形象符号、载体载能和意向意识
- 人类整个知识概念体系存在的问题
  - 以唯物论、唯心论和形式论三个基石为不同支点而形成的各种基本观点，
  - 以及在它们的基础之上构建的人类知识概念体系及其各个分支理论，都忽略了本真信息的根本地位，
  - 甚至把本真信息与载能载体、意向意识、形象符号等衍生形式之间的基本关系本末倒置。
  - 需要新的作为重构人类整个知识概念体系的理论和基础框架。
- 现有技术的缺陷或不足

- 问题：人们不能从根本上解决冗杂文本、非标形式、垃圾信息、知识爆炸和怪圈悖论等难题。
  - 对问题的深入阐述
    - 问题涉及到的主角:任何
      - 一个人
      - 一个单位
      - 一个国家
      - 一个系统
        - 包括智能网络及终端乃至独立的机器人
    - 问题存在的范围
      - 在现有的几乎任何领域的
        - 如:教育、科技、经济、政治、外交、军事、法律、医疗卫生和日常生活等领域
      - 不同方面、
        - 算、产、学、研、用方面
      - 不同阶段、
      - 不同层次
      - 不同系统，
    - 问题对主角造成的影响
      - 处于冗杂文本、垃圾信息、指数爆炸和怪圈悖论的包围之中，
      - 既搞不清楚自己所收到的这么多知识信息的本质含义，
      - 又不知道如何准确无误地发出自己应该发出的有的放矢的知识信息，
      - 也不明白自己所处的内、外、大、小环境中实际存在的各种重要的知识信息的基本内涵，
        - 包括正面的激励信息和反面的预警信息
  - 假设附加问题
    - 主角处于不利的竞争地位，或面临危机
  - 问题的后果
    - 主角的处境将会是一种糟糕的状态
    - 主角的命运将是一个艰难的过程
    - 主角被时代淘汰
- 现状：现有技术能对受限范围的知识信息及真实文本进行局部的量化处理，在非受限范围一筹莫展。
  - 协同智能的发展在知识信息数据处理领域受到了现有技术的瓶颈制约。
  - 现有软件工程及知识工程体系还缺乏从受限范围到非受限范围的转换机制，缺乏总体标准。

- 现有技术的缺陷或不足
  - 以知识信息数据处理为例
    - 直接的模数转换
    - 间接的编码转换，
  - 以金融监管（涉及证券、期货、外汇交易各方对价格信息的量化处理）为例，
    - 人们还没有找到有效的定性分析、定量分析和结构分析方法
    - 人们普遍认识不到：金融的本质是智融，缺钱的实质是缺智。
- 本发明的解决方案
  - 提供一种知识信息数据处理方法及产品，
  - 通过对全域基因文本元素的完全归纳和对已知域及目标域基因文本组合的相对完全归纳，
  - 解决知识信息的计量及测度的难题，促使人工智能与人类智能优势互补，
  - 形成效率更高的协同智能，促成形式信息革命向语义信息革命的时代飞跃。
  - 类比1
    - 信息论创始人仙农提出了形式信息数据处理技术及标准，
    - 本发明人提出了语义信息数据处理技术及标准；
  - 类比2
    - 全球定位系统（GPS）、柔性加工系统（FMS）和横断扫描仪（CT）
    - 本发明的全域数码化的网络、出版物和终端
  - 类比3
    - 人工智能、电脑和电信网获得了极大的发展，
    - 本发明要使协同智能获得前所未有的发展
- 本发明的具体任务
  - 一、提供文化基因工程方法即全域数码定位方法；
  - 二、提供全域数码定位系统，产品形式包括：
    - 1、纯形式的全域数码化网络，即知识信息数据处理领域的"GPS"；
    - 2、纯文本的全域数码化出版物，即知识信息数据处理领域的"FMS"；
    - 3、纯数码的全域数码化终端，即知识信息数据处理领域的"CT"。
- 本发明的理论依据
  - 融智概念体系
    - 是指：以义、文、物、意为基础而构成的协同智能主体的知识概念体系。
    - 量化形式：采用四元数形式。
    - 形式信息：涉及文、物，其中，
    - 语义信息：涉及义、意（注：现有的意义理论及语义信息技术没有明确区分义与意）。
    - 以曲、棋、语言为例，对上述概念及原理的基本含义说明如下：

本发明的理论依据

- 融智概念体系
  - 是指：以义、文、物、意为基础而构成的协同智能主体的知识概念体系。
    - 义，指本真信息；
    - 文、物、意统称本真信息的广义文本。
      - 文，指符号形象；
      - 物，指载体载能；
      - 意，指意识意向。
    - 所谓协同智能主体，是由人工智能和人类智能构成的新一代智能主体。
      - 其功能是对体现义的基因文本元素及组合（见：实例）进行完全归纳及相对完全归纳，
      - 以及，由此发展的知识信息数据处理（包括广义及狭义的处理）能力，
      - 具有专家系统与专家群体的整合或综合优势。
    - 范围：涉及对人类现有的整个知识概念体系的量化重构
  - 量化形式：采用四元数形式。
    - 能够通过复数域及复平面乃至曲面出入多元数系及多维空间
  - 形式信息：涉及文、物，其中，
    - 文，包括：图、文、数（注：数字信息技术属于此范围）、表、音、像等形式，
    - 物，包括：音、像、立体、活体等形式；
    - （注：音、像，涉及文与物的交集。）

- 语义信息：涉及义、意（注：现有的意义理论及语义信息技术没有明确区分义与意）。
- 以曲、棋、语言为例，对上述概念及原理的基本含义说明如下：
  - 义：曲、棋、语言的机理（含：法则）
  - 文：展示其机理的文化形式，如：乐谱、棋谱、文字或字母或动作形态等，是符号形象
  - 物：展示其机理的物化形式，如：琴、棋、传感器官（涉及使用过程）等，是载体载能，
  - 意：演奏者、下棋的人、智能主体的选择（包括以虚拟或实体的形式体现的意），是意识意向
- 信息基本定律
  - 本真信息，唯一守恒；
  - 基因文本，对应转换；
  - 基因通式，序趣简美；
  - 特式特例，非非各平（即：非对称、非同步、各自平衡）。

本发明的实现方案

- 基本方法：
  - 一种知识信息数据处理方法，
  - 相应的基本产品：
  - 相应的派生产品：
  - 进一步的派生产品：
- 详细步骤进一步综合说明如下：
- 本发明的有益效果或好处是：

本发明的实现方案

- 基本方法：
  - 一种知识信息数据处理方法，
    - 是对语义信息及真实文本进行定性、定量及结构分析的文化基因工程方法，
    - 其特征是：
      - 从广义真实文本中提取、剪接或重组文化基因，
    - 步骤是：
  - 相应的基本产品：
    - 一种知识信息数据处理产品，是由码、卡、表、库、网、端构成的全域数码定位系统，
    - 相应的基本产品的生产方法：
      - 是以码代式进行知识信息数据处理的方法，
      - 其特征是：
        - 选择纯数字形式和纯载体形式并使之相结合构成全域数码定位系统，
      - 步骤是：

- 相应的派生产品：
  - 一种知识信息数据处理产品，
  - 派生产品的生产及使用方法：
    - 是以式代码进行知识信息数据处理的方法，
    - 其特征是：
      - 使用并依托全域数码化网络的基准参照系及应对参照系，
    - 步骤是：
- 进一步的派生产品：
  - 一种知识信息数据处理产品，
  - 进一步的派生产品的生产及使用方法：
    - 是以码代式进行知识信息数据处理，
    - 其特征是：
      - 由全域数码化网络及全域数码化出版物支持的全域数码化终端，
      - 含交换机及服务器和用户计算机及其它终端装置或载体的一元基因文本终端。
    - 步骤是：
- 详细步骤进一步综合说明如下：
  - 1、确定基准参照系（化无限为有限）和应对参照系（变抽象为具体），
  - 2、以本真逻辑为序（化无序为有序），
  - 3、明确代码与载体（变失控为可控），

- 附图
  - 图1是本发明的数学模型原理图。
  - 图2是本发明的物理模型原理图。
  - 图3是本发明的技术性能一览表。
  - 图4是本发明的知识工程示意图。
  - 图5是本发明的产权监管示意图。
  - 图6是本发明的默契通信示意图。
- 本发明的有益效果或好处是：
  - 建立基准参照系及应对参照系，
  - 解决因知识信息错位而引起的冗杂文本、非标形式、垃圾信息、知识爆炸和怪圈悖论等难题。
  - 打个比方来说，上述这些难题就好比是一团乱麻。
    - 采用现有理论的范式和现有技术的办法就是：
      - 要么是快刀斩乱麻，这是从整体介入的通常作法；
      - 要么是慢慢地去理，这是从局部介入的通常作法；
      - 要么是：先把这团乱麻分成一堆一堆的小团，
    - 本发明采用的方法则是：
      - 好比是先给这团麻绳一根一根地染上色，
      - 然后，再针对具体情况（如：已知域基因文本组合）
      - 这实质上指出了解决知识信息的计量及测度难题的最佳途径，
      - 例如

- 要么是：先把这团乱麻分成一堆一堆的小团，
  - 再分别从各个小团的内、外两个方面逐步深入地去理或斩
    - （注：最后能否解决这团乱麻带来的所有问题，就要另当别论了！
    - 一方面，要看这团乱麻本身乱的性质及程度，
    - 另一方面，则要看你是否真正想要解决问题或打算把问题解决到什么程度），
  - 这是先整体、后局部，乃至不断往复逐步深入的方法。
- 本发明采用的方法则是：
  - 好比是先给这团麻绳一根一根地染上色，
    - 有的甚至再编上号（即：从全域基因文本元素的统一编号入手），
  - 然后，再针对具体情况（如：已知域基因文本组合）
    - 并考虑实际需要（如：目标域基因文本组合），
    - 既有根据又有针对性地，选择不同的颜色及编号（如：确定具体的参照系乃至参照系的组合）
    - 或分、或斩、或理（悉听尊便！）。
  - 这实质上指出了解决知识信息的计量及测度难题的最佳途径，
    - 并且，从方法及产品两方面给出了简明的示范。
  - 例如：
    - 基准参照系的应用及其效果，
    - 应对参照系的应用及其效果，
    - 由此可见，本发明的重要性、必要性和可行性，不仅易于理解而且容易证实。

例如：

- 基准参照系的应用及其效果，
  - 充分体现了人工智能主体以高速运算和海量存储支持的演绎及完全归纳优势，
  - 表现为对数据的狭义处理和广义处理的发散或收敛的精密性或专一性；
- 应对参照系的应用及其效果，
  - 则充分体现了人类智能主体以跨时空的联想、想象、灵感、直觉支持的类比及非完全归纳优势，
  - 表现为对知识或信息的狭义处理和广义处理的发散与收敛的易变性或多元性（粗放性或通用性）。
  - 有必要通过协同运行，使它们优势互补；
  - 从载体方面看，就是要克服物理载体的机械性和生物载体的易变性，即实现人机之间的取长补短。
- 由此可见，本发明的重要性、必要性和可行性，不仅易于理解而且容易证实。
  - 如果能够逐方面、逐阶段、逐层次、逐系统地实施本发明，
  - 那么知识信息的计量及测度这一国际难题的解决必将成为现实。
  - 那时，人类利用知识信息的整体水平和综合效能必将显著提高。
  - 进而，必将促使人工智能与人类智能的优势互补，形成效率更高的协同智能，
  - 促成形式信息革命向语义信息革命的时代飞跃。

- 从性能方面看，本发明，
  - 既可在信源端一劳永逸地解决非标形式的识别难题，
  - 又可通过信道实现默契通信以及借助交换终端实现一通百通的形式变换
    - （包括多种语言乃至多元基因文本形式的通译）排除冗杂文本的干扰，
  - 还可在信宿端利用一元基因文本直接走短程捷径
    - （防止或避免垃圾信息、知识爆炸和怪圈悖论），
  - 进一步，还可利用本发明对广义真实文本中涉及的知识信息进行自动测评，
    - 包括数量计算和价值评估。
  - 总之，本发明通过人工智能与人类智能、电脑与人脑、电信网与神经网的优势互补，
    - 不仅能够显著地提高协同智能主体的效率及效能，
    - 而且能够显著地改进现有技术这种在总体上仍是高消耗、低效率的知识信息处理方式。
- 实施例
  - 实施例1
    - 全域数码化网络，是由码、卡、表、库、网、端构成的融智网络（RZ-net），
    - 其特征在于：
  - 实施例2
  - 实施例3
- 结合附图对本发明作进一步说明：

实施例

- 实施例1
  - 全域数码化网络，是由码、卡、表、库、网、端构成的融智网络（RZ-net），
  - 其特征在于：
- 实施例2
  - 全域数码化出版物，
  - 是由码 + 式、卡、表、库、网、端构成的融智出版物（RZ-print for client/server），
  - 由于码与其它具体的式的对应转换或联系，
  - 整个软件系统表现为纯数字系统的子集与其相应的具体的式的子集的交集或并集，
  - 如：以（bi&...）表示的图、文（注：狭义的文）、数、表、音、像等各种具体的基因文本形式，
  - 其特征在于：

- 实施例3
  - 全域数码化终端，是由式 & 码、卡、表、库、网、端构成的融智终端（RZ-client/server），
  - 整个系统表现为多元基因文本的一元数码文本形式，
  - 其特征在于：

本发明的另一种表述（简明扼要）

（通过比喻进一步说明本发明的十要素及其相互关系与功能效果）

- 域、位、点、式、码、卡、表、库、网、端的含义及特征：
  - 域，即范围，在此是指：基因信息及基因文本元素构成的全域，
  - 位，即位置，在此是指：基因文本元素及组合在基准参照系及应对参照系中的位置或序位。
  - 点，即知识点，在此是指：基因信息元素及组合（被智能主体所认知的知识要点，如：三基一例）
  - 式，即表达式，在此是指：多元基因文本元素及组合。
  - 码，即数码，在此是指：全域数码，包括多元数及其二进制数字形式，
  - 卡，即活页卡，在此是指：单元化操作页面，
  - 表，即活动表，在此是指：多元化操作界面，
  - 库，即数据库，
  - 网，即通信网，
  - 端，即终端，
- 上述十个方面的相互关系：

- 上述十个方面的相互关系：
  - 如果把本发明的产品形式比喻为一个超级剧场，那么，
  - 域，就是该剧场所有的座位及其编号可能涉及的任何一个范围，涉及各个实在或虚拟的分剧场；
  - 位，就是该剧场的各个被实在或虚拟地设置或占用的序位；
  - 点，就是该剧场的座位及编号被实际或虚拟地占用的具体分布情形；
  - 式，就是该剧场的座位及编号被占用的情形的称谓或叫法，涉及各种称谓体系；
  - 码，就是该剧场的座位的编号，
    - 它是上述所有的称谓体系中的一种既最全面又最简捷的标准称谓体系或全域数码体系；
  - 卡，就是该剧场的最简捷的标准座位形式；
  - 表，就是该剧场的座位及编号指南或视频向导；
  - 库，是该剧场的座位编号的集中形态；
  - 网，是该剧场的座位编号的分布形态；
  - 端，是记录该剧场占用情形的装置（或记录符号），包括"出入口"。
- 二、文化基因工程产品的三个实施例：

- （1）由"号码体系"和"座位体系"构成的纯形式系统产品，
  - 它相当于该超级剧场的整个号码座位体系；
  - （2）由"实际入场"的已知域及目标域其它多元基因文本构成的纯文本系统产品；
  - （3）由"虚拟入场"的已知域及目标域全域数码文本构成的纯数码系统产品。
- 一、文化基因工程方法，其步骤是：
  - 1、在复数代码体系与二进制数字体系之间建立一一对应关系，
    - 从而，构成对指代文化基因的全域数码及全域数码组进行全域定位的数学模型。
    - 它涉及已知域及未知域，相当于一个拥有无穷多个座位及编号的超级剧场的"号码体系"；
  - 2、以卡为单元载体，在码与卡之间建立一一对应关系，用表对之进行操作，以库、网、端的形式，
    - 构成对承载基因文本的元素卡及结构卡进行全程扫描的物理模型。
    - 涉及目标域，它相当于上述超级剧场的"座位体系"；
  - 3、由码、卡、表、库、网、端，构成全域数码定位系统，
    - 如：实施例1，它相当于该超级剧场的整个号码座位体系；
    - 从真实文本中"对号入座"地自动提取相应的基因文本，
    - 从而，构成自由剪接或任意重组已知域及目标域基因文本的柔性加工用户模型，
    - 如：实施例2和3，它们相当于该超级剧场的以下"入场情形"，
    - 即：
    - "实际入场"（有"入场者"，即已知域其它多元基因文本）和
    - "虚拟入场"（无"入场者"，只有目标域其它多元基因文本的代码）两种情形]。

后续的发展和具体化，见“潜科学”

第 38 期

融智学专著及其知识要点和基本术语（一）

第 39 期

融智学专著及其知识要点和基本术语（二）

工程融智学部分：

a. *GLPS*（全球**语言**定位系统）的一个实施例

b. *GKPS*（全球**知识**定位系统）的一个实施例

第 43 期 *

融智学应用实例（*cooperating with computer e.g.*）

基于 GTCM 与 GSCM 的协同智能计算模式

GTCM(**文本**总量控制模型)的一个实施例

GSCM(**音节**总量控制模型)的一个实施例

融智学新范式

（正文的小标题）

一、融智概念体系

二、信息基本定律

三、文化基因通式

附录：

爱因斯坦成败启示

伽罗华遭遇的启示——历史的教训值得注意

我们走在获诺奖的路上——关于中国获诺奖的可能性与必要性的一些看法

系统科学专家有针对性的评语：之一　之二　之三　之四

融智学新范式[2]

引言

新范式是一种崭新的分类体系和数理形式。它是继人类智能主体与人工智能主体之后的协同智能主体的概念体系。

借助个人计算机和国际互连网，新范式产品化或工程化之后，可随时、随地、有针对性地、轻而易举地集人类知识之大成。可预测：新范式的推广普及必将导致继古希腊哲学和近现代科学之后人类认识史的第三座丰碑。

正文

古希腊哲学和近现代科学是人类认识史的两座丰碑。新范式不是哲学和科学的各门学科简单相加。三者之间是整、分、合的关系。

---

[2] 2000 年,初次公布(非正式发表)于"系统科学之窗" 论文专区(徐　辉 审稿)
2002.04 公布(正式发表)于　　(张学文审稿)
2005 年,中国谋略科学网__中国人民解放军军事统筹学会谋略研究中心科研方法智能科学
（　　）

哲学，只见“森林”。自从柏拉图提出理念论以来，唯物论、唯心论和语言论的发展，并未改变真理论众说纷纭的状况。哲学正面临学科与应用的“双重危机”。

科学，只见“树木”。自从亚里士多德创立属和种的分类体系以来，各门学科在数理论证与物理实证两方面的发展水平已相当惊人，但学科领域之间的有效交流也日益困难。不仅数学、逻辑学、语义学等领域存在数理悖论，而且物理学、化学、生物学、宇宙学、生态学、社会学、语言学、心理学、经济学、工程学、教育学等领域也存在事实悖论。个人计算机和国际互连网的普及，为推广新范式，提供了强有力的工具。　这就是新范式提出的背景。

新范式，既见“树木”又见“森林”。从信息基本定律假设提出至今，经过二十多年的理论探索和实践尝试，才形成了融智概念体系、信息基本定律和文化基因通式这一完整的新范式。它不仅为性能优于人类智能和人工智能的协同智能的发展奠定了理论基础，而且为我们提供了一种集人类知识之大成的标准架构，对基础理论、应用理论、操作技术及其工程化和商品化的全过程的知识处理十分有用。

一种真正能够随时、随地、有针对性地、轻而易举地集人类知识之大成的新范式，涉及对人类现有整个知识概念体系的量化重构。

一、融智概念体系

所谓融智概念体系，是由义、文、物、意构成的知识分类体系。在此，义，指本真信息；文，指符号形象；物，指载体载能；意，指意识意向。其中，文、物、意，统统被视为展示本真信息的广义文本。其本质特征在于对义与意进行严格区分。

通过以下事例，对上述概念的基本含义进一步说明如下：

原理，如杯子的机理，是本真信息，属于义的范畴；展示其机理的文化形式，如杯子的图纸，属于文的范畴；展示其机理的物化形式，如具体的杯子，属于物的范畴；智能主体的选择，如杯子的构造及外观的设计构想，属于意的范畴。

可以说，广义真实文本涉及：图、文、数、表、音、像、立体、活体等具体形式，本真信息则涉及：广义真实文本存在和变化的机理。

新、旧范式的区别在于：

旧范式，对义与意不作区分。例如：现有的语义学、意义理论和语义信息理论，不仅都没有明确地区分义与意，而且，总是用意义的概念把义与意混为一谈。

新范式，对义与意严格区分。例如：《一种知识信息数据处理方法及产品》的原理，不仅对义与意作了严格的区分，而且对义、文、物、意作了明确的定义。

如果说文与物涉及图、文、数、表、音、像、立体、活体等形式信息，那么，义与意则涉及义、意、意义等语义信息。个人计算机革命和通信革命的成果主要体现在形式信息处理方面，其困难则主要发生在语义信息处理方面。

试问：旧范式在语义概念本身都存在问题的情况下，如何解决语义信息领域面临的难题？又如何解决语义信息和形式信息交错的难题？何况形式信息领域本身面临的难题并没有彻底解决。

新范式，则把义、文、物、意四个分类系列作为协同智能主体进行定性分析的分类基础，区别于现有范式及其分类体系。

二、信息基本定律

所谓信息基本定律，即：本真信息，唯一守恒；广义文本，对应转换；基因通式，序趣简美；特式特例，非非各平（即：非对称、非同步、各自平衡）。

其中，只有本真信息才是唯一且守恒的，广义文本则必然是多元且冗余的。

广义文本，主要是指展示文化基因的子全域和超子域；对应转换，则是指广义文本的相互转换遵循同义并列法则。

所谓子全域，是由完全归纳的基因文本元素构成的全域的各个一元文本子集。例如：由 26 个字母构成的英语字母表，就是英语这种文化形式的基因文本元素的子全域。又如：由 A、T、C、G 四种核酸或四个硷基构成的基因密码表，就是生物这种物化形式的基因文本元素的子全域。

所谓超子域，是由非全归纳的基因文本组合构成的各个一元或多元文本子集，即：在时空变换方面，能够超越子全集。 例如：由字母的线性排列构成的英语词素、词、词组、句子、段落、篇章等，则是英语这种文化形式的基因文本元素组合的超子域的各个进化阶梯的不同发展形态。又如：由核酸或硷基排列组合构成的氨基酸、蛋白质、染色体、细胞、组织、器官、系统、机体等，则是生物这种物化形式的基因文本元素组合的超子域。

基因通式，是指包含所有子全域的全域多元数通式；序趣简美，是指全域及其子全域的构成法则。特式特例，是指各个子全域和超子域；非非各平，是指各个特式或特例相互之间具有的空间上的非对称性和时间上的非同步性，以及它们的平衡状态与趋动转换过程

三、文化基因通式

所谓文化基因通式，是指表示义、文、物、意体系及定律的标准代码形式，即全域多元数通式，（$a+bi\&\cdots$）是表示本真信息和广义文本的最简单、最完备的符号形式，复数系（$a+bi$）和四元数系（$a+bi+cj+dk$）等都是它的子系统。

$t=a_0+\{a_1, a_2, \cdots\}$的几何形式是一维数轴，（$x, y, z$）的几何形式是三维坐标，四元数系的几何形式是（$x, y, z, t$），多元数系的几何形式是（$x, y, z, ict$[3]）。程序分析依据一维参照系进行，结构分析依据三维参照系进行，序位分析依据多维参照系进行。

结语

[3] 或“$t_1\ t_2\ \ldots\ t_n$”，用于表示**自然界**（**人工界**和**心理上的**时钟均为其特例）**一系列“时钟体”的时间维**。

全域数码定位系统，涉及协同时空观，即：由具体的一维线性子集和三维立体子集共同构成的动态协同序位分析框架。其中，图、文、数、表、音、像、立体、活体等任何形式的基因文本元素及组合（包括子全域和超子域）都有各自相应的序位。

协同智能的原理，就是人类智能与人工智能在该框架体系内形成的优势互补机制。

文化基因通式（*a+bi&…*）是全域数码化的定量分析工具。复数域及复平面是出入多元数系及多维时空的转换界面。

新范式，在理论上以融智学命名，在实践上以文化基因工程落实。

融智学的基本含义在于促进哲学和科学的各门学科的融通，为协同智能的实现提供崭新的知识概念体系。

协同智能主体，是指由人工智能和人类智能的优势互补构成的新一代智能主体。其基本功能是对体现本真信息的基因文本元素及组合（包括程序、结构以及框架等）进行完全归纳或相对完全归纳，以此发展出只有专家系统与专家群体充分协作才会具有的知识信息数据处理（包括广义及狭义的处理）的整合功能。它具有的协同时空观、完全归纳逻辑、全域序位参照系，是现有的人工智能和人类智能两类主体各自都不具备的。

总之，在哲学和科学的各门学科的旧范式中存在的一系列根本性的难题，在融智学新范式和文化基因工程及其赖以发展的协同智能主体面前，都将迎刃而解。

附录：

爱因斯坦成败启示（一）

（根据国际上公开的权威数据摘编）

爱因斯坦于 1905 年首次提出狭义相对论原理，论文发表在《物理学年鉴》上，同年，他对狭义相对论作了重要补充，并为辐射问题建立了最初形式的质能关系式。1907 年，爱因斯坦完成了一篇通俗性的相对论文，其中包含一般形式的质能关系式：E＝mc2。他的卓越论文建立了全新的质量、时间和空间概念，并向同时性观念提出了挑战。相对论的伟大意义在于：它抛弃了“绝对”时空观和空间充满以太的思想；当时，以太被看作是光以及其它形式的电磁波传播媒介。现在看来，１９０５年６月爱因斯坦关于相对论的开创性论文在《物理学年鉴》上发表，是理论革命阶段的典型例子。

玻恩１９０６年在哥廷根研究“运动物体的电动力学和光学”时，竟然还未听说过爱因斯坦和他的工作。１９０６——１９０７年间，英国剑桥大学的情况也是如此。

按照爱因斯坦的妹妹的回忆，爱因斯坦当时“想象在有名望的、拥有众多读者的杂志上发表论文，是立即会引起注意的”。当然，他期望“强烈的反对和最严厉的批评”，但缺少反响和“冷酷的批评”反而使他“非常失望”。不久，他收到普朗克寄来的一封信，就论文中几个模糊不清之点提出问题，这使爱因斯坦感到“异乎寻常的高兴”，因为普朗克是“当时最伟大的物理学家之一。相对论后来迅速变成了物理学家们感兴趣的讨论和研究课题，这种戏剧性转变主要由于普朗克较早较深地介入相对论所引起的。

爱因斯坦论文发表的第二年，普朗克就开始在柏林讲授相对性理论，但不是爱因斯坦理论内

容，而是洛伦兹的电子理论。1907 年，普朗克的助手劳厄（后来的诺贝尔物理学奖获得者）发表了一篇关于相对论的专题文章。

1906 年 6 月，普朗克在德国物理学会上发表关于相对论的演讲（同年刊登在杂志上）；1907 年，在普朗克的指导下，莫森格尔完成了第一篇专论相对论的博士论文。佩斯指出，早期介入这一领域的人实在是太少了，乌尔茨堡的劳布和拉登伯格是为数不多的几个例外。劳厄曾经来到伯尔尼拜访爱因斯坦，他发现很难相信这个“年轻人”竟然是“相对论之父”。几年后，劳厄撰写了一篇非常出色的介绍相对论的学术论文。劳厄在 1917 年 3 月 24 日写给爱因斯坦的信中，表达了他对自己的革命性工作的兴奋之情：“终于实现了！我的关于波动光学的革命观点发表了”。他接着写道：在“这一紧要关头”，它们“无疑会激起每一个爱好和平的物理学家最强烈的憎恨”；但“我仍然要坚持这些备受谴责的观点”。

除了玻恩自己介绍了他是怎样第一次听说相对论的之外，我们还从英费尔德那儿了解到当时的一些情况。英费尔德曾谈到他的朋友洛利亚教授告诉他的一件事，洛利亚的老师“克拉克夫大学的维特柯夫斯基教授（他是一个非常伟大的教师）”读了爱因斯坦 1905 年关于相对论的论文后，“冲着洛利亚兴奋地喊到”：“读读爱因斯坦的论文吧，又一个哥白尼诞生了！”又过了一段时间（玻恩说是 1907 后）洛利亚在一次物理学会议上遇到了玻恩，他向玻恩谈起爱因斯坦，并问他是否读过那篇相对论论文。结果，“不光是玻恩，在场的每一位都从未听说过爱因斯坦。”英费尔德说，“玻恩立即认识到相对论的伟大，同时感到有必要对它作形式上的推广。英费尔德认为，玻恩后来对相对论的研究工作，“是早期对这一科学领域作出的重要贡献。”

最初，表示愿意接受爱因斯坦狭义相对论的物理学家很少，因此不足以在世界范围内开创一场科学革命，但来自德籍理论物理学家的反对意见却很多。1907 年 7 月，普朗克在致爱因斯坦的信中说：“相对性原理的倡导者”仅仅形成了”小小的一个圈”，由此他认为，他们之间“取得一致意见倍加重要。”“相对性原理”既体现了普朗克个人偏爱的洛伦兹理论，也体现了爱因斯坦的相对论。然而，爱因斯坦的声望在持续增长，尽管仍相当缓慢。1907 年秋，斯塔克（《放射性和电学年鉴》的编者）写信给爱因斯坦，要求他写一篇相对论的“评述文章”。

一篇引用爱因斯坦相对论论文的文章是考夫曼 1905 年写的。他认为爱因斯坦的“研究与洛伦兹的研究在形式上是同一的，”只不过是后者的有益于推广。考夫曼最后说，他自己的实验数据驳倒了爱因斯坦和洛伦兹的电子理论。过一会儿，我们将回过来研究这个问题。

1907 年，爱伦菲斯特写了一篇以爱因斯坦理论为主题的论文。第二年（1908 年），闵科夫斯发表文章，把爱因斯坦理论根本性地转化为数学形式，“大大简化了狭义相对论”。经过这样几个步骤，纸面上的革命才变成了真正的科学革命。佩斯指出，从 1908 年开始，爱因斯坦的名声和他的影响都在迅速提高。

此时，爱因斯坦这颗科学界明星升起来了。1909 年春，他从伯尔尼瑞士专利局一个地位低微的专利审查员，一跃而成为苏黎世大学理论物理学助理教授，这很明显是由于他在固体量子论方面所做的工作。爱因斯坦的推荐人之一写道：爱因斯坦“当属最伟大的理论物理学家之列”，“由于相对论原理方面的工作，他正受到极其广泛的重视”。1909 年 7 月 8 日，爱因斯坦获得了日内瓦大学的荣誉学位，同时获得了这项荣誉的还有奥斯瓦尔德和居里夫人。他在这个职位只呆了两年，1911 年 3 月他又来到布拉格，晋升为德国卡尔—费迪南大学正教授。在那儿工作了 16 个月后，弗兰克接替了这个职位。爱因斯坦又返回苏黎世，担任综合技术学院物理学教授。

当然，影响接受狭义相对论的困难主要是概念上的，但也确实存在实验上的障碍。在 1905 年开创性的论文末尾，爱因斯坦推导出了一个电子横向质量公式。这个公式与洛伦兹理论中的公式极其相似，其中的差导很快就被消除了。于是，这样两种理论能给出相同的结果。但是，考夫曼在分别发表于 1902 年和 1903 年的论文中指出，他的实验结果与洛伦兹理论（同样也适用于爱因斯坦理论）的预言有很大差异。爱因斯坦对这些结果无动于衷。1906 年，考夫曼在《物理学年鉴》（一年前爱因斯坦发表相对论论文的同一杂志）发表了一篇文章，详细归纳了爱因斯坦的时

空观念，谈到了洛伦兹—爱因斯坦的电子理论。他总结说，他自己的测量结果与洛伦兹—爱因斯坦理论的“基本假设是不相容的。”洛伦兹因此写了一封信给彭加勒，说他自己已经走上了“末路”。他对彭加勒说，“不幸的是”，他的假说“与考夫曼的新实验矛盾”，他认为“不得不放弃它”。但爱因斯坦却深信：实验数据与理论间“系统误差”的存在说明有“未被注意的误差源”；新的更精确的实验一定会证实相对论理论。爱因斯坦的话得到了证实，1908 年布歇尔发表了新的实验结果，完全符合洛伦兹和爱因斯坦的预言。1910 年，胡普卡的实验对此再次予以确证。而决定性的结果是 1914—1916 年获得的。从那以后，各种表明相对论正确性的论据不断出现，且极为丰富。

随着实验证据的实现，相对论本身进行了根本性重构。这项工作是哥廷根大学数学教授闵科夫斯基完成的。有趣的是，几年前，闵科夫斯基教过爱因斯坦数学。1908 年，闵科夫斯基发表论文，引进四维“时空”概念，取代了孤立的三维空间外加一维时间的不相容概念，他还把相对论转化为现代张量形式（这要求物理学家们进一步学习由里奇和列维—西维塔建立的新数学理论），在相对论中引进专业术语，并明确指出：以相对论观点看，传统的牛顿引力理论已经不够用了。很明显，爱因斯坦开始并没有理解闵科夫斯基工作的意义，甚至认为把他的理论改写成张量形式是“多余的技巧”。但到了 1912 年，爱因斯坦终于转变过来了；1916 年，他以感激的心情承认闵科夫斯基使他大大地简化了从狭义相对论到广义相对论的过渡。爱因斯坦后来着重强调了闵科夫斯基的贡献，他说：如果没有他，“广义相对论也许还在襁褓中。”

闵科夫斯基的时空观首次公开发表是在 1907 年 11 月 5 日的一次演讲中，演讲的标题是“相对论原理”。但这篇演讲直到闵科夫斯基去世后 6 年的 1915 年才出版。不过庆幸的是，早在 1908 年和 1909 年发表的两篇论文，闵科夫斯基已阐述了他的时空观思想。闵科夫斯基充分认识到他的贡献的重要性，他在 1907 年演讲时开宗明义地说：“先生们，我想向诸位讲述的空间和时间观念……是根本性的，……因此，孤立的空间和时间本身将注定消失在阴影里。”事实上，闵科夫斯基在这篇演讲的第一稿上，把他的新时空观的“特征”说成是“革命的”，而且是“极端革命的。”可是，在演讲稿最后付印时，“革命的”词语省去了。

玻恩向我们讲述他最初阅读爱因斯坦论文的经过，使我们懂得爱因斯坦的概念是多么深奥难懂，甚至对那些异常精通数学的人亦是如此。1907 年，当洛利亚向他们介绍爱因斯坦论文的时候，玻恩正是闵科夫斯基的大学研究班成员，因此“对相对性思想和洛伦兹变换很熟悉。”他回忆说，即便如此，在阅读爱因斯论文时，“爱因斯坦的推理超出我的意料之外。”玻恩发现，“爱因斯坦理论是全新的和革命的”，是天才的创造。爱因斯坦的观点“向牛顿建立的自然哲学以及传统时空观大胆提出了挑战。”现在看来，玻恩确实认识到了爱因斯坦的思想革命和理论革命的威力，但也清醒地看到真正的科学革命尚未到来。玻恩戏剧性地指出，事实上，爱因斯坦理论是如此激进（亦即新的和革命的），以至必须“做出努力才能很好地予以消化和吸收。”而且他还提醒我们，“并不是每一个人都能够或愿意这么做。”看来当初他本人是做到了。爱因斯坦革命要求人们普遍接受爱因斯坦关于物质世界的全新思考方式。

1909 年美国科学家刘易斯和托尔曼发表的文章，清楚地说明了接受爱因斯坦假说的实际困难。他们承认爱因斯坦相对论原理“综合了大量实验事实，没有出现矛盾的反例，”其中他们列举布歇尔的实验作为支持这一理论的重要依据。然而，他们感到相对论基本“原理”的这一方面无可挑剔时，另一方面就暴露出了问题。例如，在对“绝对运动无法观察到”这一普遍原理表示理解时，他们却发现相对于任意独立观察者光速不变相对性“奇异结论”，这可能是“基于某种感官心理学上的科学幻想”。

时间一年年地过去，越来越多的物理学家终于转变了。然而，他们当中有许多人只接受爱因斯坦公式，承认“收缩性”是光速不变性引起的空间问题的基础。但仍然坚持绝对时间和同时性的信仰（包括洛伦兹在内）。1911 年 4 月，法国物理学家朗之万在哲学家大会上发表演说，为相对论增添了更轰动性的因素。朗之万是位卓越的科学家，爱因斯坦曾经说过，如果他没有发现狭

义相对论，朗之万将会发现。在讨论时间相对性或膨胀问题时，朗之万没有采用爱因斯坦那种利用运动时钟和静止时钟解释时间效应的费解作法，而是用所谓的“双生子悖论”取代了爱因斯坦的“时钟悖论”，并立即成为大众所熟知的由相对论引起的怪物。相对论时间问题是这样产生的：如果双胞胎兄弟一个留在地球，另一个去星际空间旅行，那么当旅行的兄弟返回地球时，竟会发现双胞胎的年龄不同了。朗之万列举的一个例子是，旅行者直线飞向一颗恒星，绕它一周并原路返回。如果旅行的速度足够大（当然比光速小），最后旅行者将发现，在他两年的旅行中，地球已经度过了漫长的两个世纪。哲学家亨利•伯格森后来承认，正是朗之万 1911 年 4 月的演讲，“第一次唤起了我对爱因斯坦观念的注意”。

时钟（或双生子）悖论很快成为（在某种程度上今天依然）相对论使人困惑或招致敌意的原因。劳厄曾谈到那些反对相对论的“思想内容”、基本公式或数学结果的人。1911 年他写信告诉爱因斯坦，反对相对论的共同理由“主要是时间相对性和由此产生的悖论”。劳厄在 1912 年写的世界第一部相对论教科书中指出：这些悖论和其它有关时间相对性问题具有“伟大的哲学意义”，“正是由于这一原因”，“只能用哲学方法”对待这些问题。我们还注意到，1911 年爱因斯坦在讨论这一见解时，使用了理想实验的方法。他假设把装有“小生物的盒子”送向“遥远的飞行里程”，结果在它返回地球时，“盒子的内部情况几乎没有变化”，而留在地球上的生物已“繁衍生息了许多代了。”

尽管许多人不愿轻易接受爱因斯坦对物理学基本思想进行彻底重构，但他们却已在应用爱因斯坦的数学结果。劳厄（以及其它人）曾指出，这些数学结果在形式上和洛伦兹理论结果是一致的，但它们的“物理本质”却有重大差异。劳厄甚至宣称（1911），两种理论的“实质差别是不可能的”。但人们很快就认识到爱因斯坦理论更加优越，特别是在广义相对论建立之后，狭义相对论的重要性尤其显示出来。

大约到 1911 年，爱因斯坦狭义相对论已有了相当数量的拥护者，导致了一场科学革命。同一年，索末菲宣布，相对论理论“已经完整地建立起来，它不再是物理学的前沿了。”1912 年初，1911 年度诺贝尔物理学奖获得者维恩建议：授予爱因斯坦和普朗克这项最高奖赏。他在推荐书上写道：从“逻辑的观点看，”相对论原理“应当被看作理论物理学所取得的最重要的成就之一。”他说，目前已有“实验明确证实了这一理论。”他作结论，洛伦兹是发现相对论原理数学内容”的第一人。而爱因斯坦则“成功地把相对论简化为一个简单的原理”。

伽罗华遭遇的启示——历史的教训值得注意

（根据国际上公开的权威数据摘编）

伽罗华获得的非凡成果，在他去世后 11 年才开始得到数学界的承认；期间风云变幻，人事繁杂，而能够从中带着巨大的荣耀脱颖而出的只有年轻的伽罗华一人。

他写出了将成为他最著名的论文（关于方程可根式求解的条件），并于 1831 年 1 月递交科学院。递交这篇论文是他最后一次尝试让数学界承认他的工作。到 3 月，科学院方面仍杳无音讯，于是他写信给院长打听他的文章的下落，结果又石沉大海。7 月 4 日，他终于打听到他给科学院的那篇论文的命运：因“无法理解”而遭拒绝，审稿人泊松是这样结束其评审报告的：“我已尽了一切努力去理解伽罗华的证明，他的推理不够清晰，不够充分，我们无法判断其正确性；本报告也不能就此提出任何想法。作者宣称，该文研究的特殊对象是具有众多应用的一种更普遍的理论的一个组成部分。也许，整个理论的各个不同的部分能相互澄清，因而比孤立的部分更容易掌握。我们不妨建议作者发表其完整的结果，以便得出明确的意见。但就目前他送交科学院的部分结果而言，我们不能推荐说应给与承认。”

站在恩赐立场上提否决意见的鉴定人，可能感到他的这个报告无可挑剔，我们也不知道他对

伽罗华其后的行为有无影响。

伽罗华写了封长信给他的朋友舍瓦利耶,其中大致描述了他的数学理论，从而给数学界留下了唯一一份它将蒙受何等损失的提要。

他遭拒绝的是篇什么样的论文呢？1843 年 7 月 4 日，刘维尔在法国科学院演说的开场白这样说：“我希望我的宣告能引起科学院的兴趣；在埃瓦里斯特•伽罗华的那些文章中，我已经发现如下漂亮的问题的一个既精确又深刻的解答：……是否根式可解？……”

伽罗华留给世界的最核心的概念是群，这对所有时代都是最有意义的概念之一，在许多数学领域有它的应用，而且可用于物理、化学和工程学分支。

这是个完全抽象的概念。他之所以有如此威力，原因是有大量群的实例存在，它们往往各具不同的特性。群的概念具有多面性，所以可用多种方式介绍它。

我们走在获诺奖的路上

——关于中国获诺奖的可能性与必要性的一些看法

2000 年 6 月中旬写于筑

获诺贝尔奖，不仅被世界公认是科学上的最高成就，而且对获奖者所在国也具重大意义。

中国想获诺贝尔奖的愿望虽然 由来已久，但是，像今天这样大张旗鼓地向世界谈及自己的诺奖之梦，对中国人来说，似乎还是第一次。

也许是因为认识到了获诺奖对一个发展中国家的特殊意义，也许是因为已增强了获诺奖的实力和勇气，21 世纪伊始，中央电视台报导了我国中青年学科带头人与六位前诺奖获得者之间的对话，前不久(2000 年 6 月初)，又在北京军事博物馆举办了一次声势浩大的诺奖展览，同时，专家学者们也纷纷议论中国人可能获诺奖的学科领域,有人估计中国获诺奖的时间是 2030 至 2050 年，有人估计是 2005 年至 2020 年，总之，大家都认为该是中国获诺奖的时候了。朱总理曾不止一次地在公开场合说中国应该获诺贝尔奖。欧美 18 个机构制订的预测 2025 年科技发展及其对公司影响的计划指出：一个主要问题是如何衡量信息和知识的经济价值。提出这方面有效理论的学者将被授予诺贝尔奖。

为什么解决这个问题可以达到获诺奖的学术水平呢？

简单地说：这个问题是由形式信息革命阶段过渡到语义信息革命阶段的卡脖子的难题。

如果能解决这个问题，那么，不仅能够顺利地实现科学、技术、文化、产业等范式的新旧转换，而且也能够更容易地实现产业升级，如加快整个世界的工业化及信息化进程。

据悉：目前中国已有人在信息科学、信息技术、信息产业和信息经济学等领域找到了解决上述难题的钥匙。按照科学革命必须的“思想、口头、笔头、认可”四个阶段的进程来看，上述成果正处于逐步扩大认可的阶段。从该成果涉及的领域和它可能带来的积极效果这方面来看，就可以知道中国获诺奖的可能性与必要性。

首先，在以下学科领域的突破会显著地提高人类的认识水平。

哲学领域，在本体论、认识论、语言论三次大的发展或转向之后，对真理论、目的论、方法论进行了革命性的改造，以《融智学》的本真信息观化解了“哲学的双重危机”。

物理学领域，在牛顿的绝对时空观和爱因斯坦的相对时空观之后，发现了更具普遍性的《融智学》协同时空观，并且，更好地阐述了“质、能、信”关系。

逻辑学领域，在传统和现代逻辑学体系之后，提出了更具普遍性的《融智学》全域逻辑体系。

语言学领域，在自然语言学、计算语言学和计算机程序等人工语言学现有成果的基础之上，通过一定的数学形式作为中介，建立自然语言与机器语言之间的对应转换关系，以多元数系规范

了“万码（马）奔腾”的软件业。

符号学及语义学领域，在指称、表达式和意义（即：“语义三角”）研究成果的基础之上，对意义理论实施变革，解决了对知识信息进行人机协同的定性、定量和结构等分析难题。

心理学领域，在人类智能和人工智能研究成果的基础之上，着重从识别、理解、表达或再现几方面探求智能概念的本质，并且，明确地提出了协同智能是人类智能及人工智能发展的下一个进化阶梯。

教育学领域，借助上述各学科领域的有关成果，在已有知识成果的基础之上提炼出文化基因元素及其组合的基因文本原型，从而发现了传递形式信息和语义信息的捷径。

经济学领域，从市场经济、法制经济、网络经济、知识经济、全面生活质量原理、生态经济等多个角度，提出了有效的信息经济学基础理论——全域数码定位理论及其方法。

然后，在科技推广方面，对相应的产业进行知识信息化改造或拓展。

计算机领域，简化形式信息的处理，并形成统一的国际标准软件体系——全域数码定位体系，即：区别于现有的各种操作系统及其应用系统和知识信息数据库的融智新体系。

通信领域，简化各种具体知识信息传递方式，以默契通信的方式最大限度地提高通信效率及效能。

传感及传播领域，简化形式信息的输入和输出，消除人机界面知识信息数据交流的各种障碍。

人工智能领域，在专家系统和领域专家的工作基础之上，构成协同智能主体及网络，在全球网上实施知识系统工程和知识产权监管，自动测评网上交流的知识信息的价值。

新闻出版及广播电视领域，借助文化基因工程减少乃至消除冗杂文本，合理配置电子出版物与印刷出版物所占用的资源。

教学培训领域，借助人工智能和文化基因工程等方面的新技术及新设备，提高人力资源的开发效率及效能。

传统产业改造领域，对农业、工业、商业和服务业进行信息化改造，显著降低其生产、管理、经营等方面的成本，使之更加精细化，进而，使其产品更容易适销对路。

综上所述，可见前述问题的解决是科学、技术、产业等新旧范式转换的关键。如果我国能在这一涉及语义信息革命的关键领域获诺奖，那么，我们的整体国力将从根本上得到显著的提升。

为此，有必要分析中国人解决这个问题的基本优势或潜在的机会。

根据中科院有关院所和中国有关名牌大学的院所以及部分顶级高科技公司的研发部门的工作现状，对近期的未来可做以下预言或判断：计算机理解语义信息的难题仍将是世界性的难题，这方面中文与西文信息处理面临的是同样的难题，目前，从理论到技术，谁都没有实质性的突破成果公开。但是，就自然人理解语义信息而言，汉语的“语义句法”似乎应比英语等“语形句法”更有优势。例如：北大徐通锵教授对“语义型语言”的研究、北邮钟义信教授对“全信息”的表述、北京语言文化大学张普教授的“动态语言知识更新研究”，中科院黄曾阳教授的“概念层次网络理论”的构想等成果，都体现了中国人在研究语义信息方面的优势。比较而言，英国佛雷格对“语言形式”的研究，瑞士索绪尔对“共时态”语言结构的探讨、德国胡塞尔对“意向”作用的强调、美国乔姆斯基对“转换语法”的深入和美国仙农对“离散信息”的定量分析等成果，则体现了欧美人士在研究形式信息方面的优势。这也许是为什么形式信息革命由说印欧语系语言的民族发起并推动，而语义信息革命则将由说汉藏语系语言的民族发起并推动的原因之所在吧！

必须指出：在形式信息主导的人类知识概念体系的总体参照系框架内，旧的科技及产业范式已发展到了登峰造极的地步。据悉：中华人民共和国国家知识产权局的档案中已多了一份新的科技及产业范式的基本发明专利文献，它是由中国人提出的语义信息主导的人类知识概念体系的新的总体参照系框架的雏形。如果我国能在国际专利的优先权保护期内组织足够的力量完善整个基本专利及外围专利体系，并及时申请国际专利，那么，既能加速中国获诺奖的进程，又会增加中国获奖的学科数量，还可使中国在新旧科技及产业范式转换之机获得彻底打一个翻身仗的机会。

“没人知道有什么样的新观念、新发明或新技术应用在哪个角落，在过去，这些发展对世界经济产生了深远的影响或改变”。特别是信息科学、信息技术以及信息经济中，那种能使一个自然人或法人乃至一个国家“一步领先步步领先”的千载难逢的机会，往往可使后进者赶超先进者。这也许是发展中国家或地区赶超发达国或地区的唯一捷径，特别是对中国这样在大部份科技及产业领域至少落后美国几十年的发展中国家提供这种机会或捷径更是难能可贵。因为，在新、旧科技及产业范式或体系转换的关键期，如果中国能够抓住这一可打翻身仗的机会，那么，一旦进入新体系或新范式，在信息科技及产业方面占据制高点之后，就更将会如鱼得水，因此，必然能获得前所未有的更大发展空间。

新、旧范式或体系的一个显著区别是：**前者，资源决定一切**，即**物质和能源**的多少决定**机会**的多少；**后者，智能决定一切**，即**知识和信息**的多少决定**机会**的多少。

根据“**金三角对策原理**”来看，**前者**是“机会**限制**愿望和能力”，**后者**是“**愿望**和**能力**制造**机会**”。即：

痴人对机会视而不见，
常人只知道等待机会，
智者会积极寻找机会，
天才则努力创造机会。

（1989 年春邹晓辉于深圳科图）

国家及其政府和企事业单位都是由无数的个人组成的。一个人的命运能否改变主要决定于其知识经验结构涉及的文化基因的可重组程度。试想一下，如果一个发展中国家或地区能有一批达到诺奖获得者水平的人才，那么它还会一直落后吗？显然，这个国家或地区再也不仅仅是一个普通的发展中国家或地区了，它一定会有较大的发展。这就好比一个语文、数学、英语三科成绩都非常优秀的中学生，他或她的其它各科成绩绝不会差到哪里去。因为，各科知识之间是有内在联系的，只要通过基础学科确立了重组文化基因的基本能力，虽然采用的方法不同，有的甚至还可能很一般，但只要经过一定的量的积累，达到了能使现有的文化基因组合产生质的变化的程度，那么就一定能够向其它学科迁移这种重组文化基因的能力。

由此可见，语义信息处理或知识定量分析，一旦达成共识并被视为发展中国家和地区打翻身仗的核心学科，发起获诺奖攻坚战的有关单位就应首先攻下这一中国人最具相对优势的关键学科领域。

系统科学专家（四川大学 陈雨思）有针对性的评语之一

知识信息处理研究的进展与前景

1.关于钟义信的全信息理论

2.关于**邹晓辉**的**融智学**理论

在知识信息处理方面, 另一个值得重视的工作是邹晓辉的融智学新范式。

从理论创新的角度 ， 邹晓辉提出人工智能、人类智能和协同智能关系的理论；义、文、物、意构成的知识分类体系；信息的基本定律（本真信息，唯一守恒；广义文本，对应转换；基因通式，序趣简美；特式特例，非非各平［ 即：非对称、非同步、各自平衡 ］）和文化基因通式等，形成了融智学新范式。我以为，融智学新范式是创立者长期研究、深思熟虑、融汇贯通的结果，是应该加以足够重视的。

基于融智学新范式而得的“一种知识信息数据处理方法及产品”是一项获得国家专利的技术，这里不准备对其技术细节做出评价 ， 而仅从理论的角度加以讨论。

为了了解融智学新范式有什么创新意义 ， 我们不妨将其与钟义信的全信息理论做一比较。

所谓融智概念体系，是指以义、文、物、意为基础而构成的协同智能主体的知识概念体系。

这与钟义信的全信息理论既是相通的 ， 也是不一样的。

邹晓辉说， 形式信息涉及文、物；语义信息涉及义、意（注：现有的意义理论及语义信息技术没有明确区分义与意）。这里的形式信息 对应于钟义信的语法信息； 这里的语义信息对应于钟义信的语义信息;协同智能 对应于钟义信的 语用信息 ， 这是融智概念体系与钟义信的全信息理论相通之处。

值得引起注意的是,融智概念体系把形式信息分为文、物；语义信息分为义、意；(语用信息归入协同智能?)。邹晓辉还特别强调：现有的意义理论及语义信息技术没有明确区分义与意。这种分法的创新意义何在呢？下面我们以信息三象理论为基点来做一分析。

前面我们指出 ， 全信息理论是以“信宿”为基点的 ， 由于信宿千差万别 ， 基于信宿的语义信息和语用信息的描述非常困难。另外我们还指出 ， 信宿的结构和结构变化表现出的功用信息 ，包含了信源、信道和信宿的结构和结构变化的信息。

因此,语义信息和语用信息的描述困难在于,1. 信宿千差万别； 2. 信源、信道和信宿的信息混杂在一起。而解决之法是 ,1. 对千差万别的信宿给以适当描述； 2. 把信源、信道和信宿的信息做分别描述。

融智学分类体系恰好对这两个问题给出了自己的回答。

首先是人机分离。即采用人工智能与人类智能既分工又协调的协同智能模式 ， 把不确定程度较大部分交由人类智能去完成;把不确定程度较小部分交由人工智能去完成 ， 然后两者协调形成协同智能。

其次是信道和信源、信宿分离,即文、物与义、意分离。融智学新范式对形式信息的理解与全信息理论是有差别的。邹晓辉指出，文，指符号形象；物，指载体载能 ， 他还用 曲、棋、语言来做比喻 ， 指出： 乐谱、棋谱、字或字母或动作等，是符号形象，即文；琴、棋、传感器官（含使用过程）等，是载体载能，即物；这从广义来讲，是对应的信道信息。

另外是信源和信宿分离,即义与意分离。邹晓辉指出，义，指本真信息；如曲、棋、语言的机理（含：法则），是本真信息，即义；意，指意识意向。如演奏者、下棋的人、智能主体的选择（包括以虚拟或实体的形式体现的意），是意识意向，即意。所谓本真信息、机理 ， 应当主要是表达信源的结构和结构变化 （ 或信源的运动状态和方式及其变化)的信息；而所谓意识意向,应该是信息选择者的意识意向,应属于信源(亦可属目标系统,此处不讨论)。

将义与意分离,或将信源和信宿分离,其重要之处在于守恒信息的突显。邹晓辉强调指出，本真信息,唯一守恒。也可以说,表达信源的结构和结构变化 （ 或信源的运动状态和方式及其变化 ）的信息在信息过程中是守恒的。

上述讨论是就单一信息传递过程而言的 ， 单一信息传递过程的复杂叠加,就形成 信息的自组织过程,而人类文化是信息的自组织过程的结果。既然单一信息传递过程存在守恒信息,那么复杂信息传递过程也必然存在守恒信息， 信息的自组织过程也必然存在守恒信息 ， 最后归结为人类文化必然存在守恒信息 ， 这就导致了文化基因的结论。

文化基因的结论意谓着人类文化均由文化基因组合而成 ， 任何文化过程都表现为文化基因的提取、剪接或重组的过程，将这个过程在电脑中实现,就是文化基因工程。

由上可见 ， 融智学新范式在理论上是有突破的,在知识信息处理问题上的价值是显然的，其市场价值也是很大的。

（引自：四川大学陈雨思“**信息自组织与知识信息处理——与邹晓辉交流**”一文）

系统科学专家（四川大学 陈雨思）有针对性的评语之二

破译文化基因 清理知识乱麻 打造信息富翁

——邹氏知识信息数据处理技术及产品介绍[4]

[4] 本文是陈雨思应邀为邹晓辉携其发明“一种知识信息数据处理方法及产品”参加“中国专利博览会”即兴而作。

计算机与因特网技术的迅猛发展，突破了信息交流的时空障碍，每个人都可以根据需要去选择网站、选择信息，也可以任由鼠标顺着自己情绪的波动而游走。人们在巨大的信息海洋中享用着无穷的信息资源。

然而，信息海洋茫茫无际，无穷的虚拟时空消耗着我们有限的时间和精力，我们怎样才能找到有用的信息呢？怎样才能高效率地组织这些信息呢？

邹晓辉先生指出“现有技术只能对受限范围的知识信息及真实文本进行局部的量化处理，所以，不能从根本上解决冗杂文本、非标形式、垃圾信息、知识爆炸和怪圈悖论等难题。”

无边无际的冗杂信息不仅消耗着我们有限的时间和精力，而且干扰和破坏我们原来有序的信息、知识和思维结构，使我们出现信息、知识和思维结构的混乱。

冗杂信息的掠夺、干扰使我们原来拥有的信息和知识的价值趋于零；茫茫无际的信息海洋使有用信息的寻找十分困难；思维的混沌使我们不知如何选择有价值的新信息； 信息的复杂使我们不知如何高效率地组织信息，于是我们成了信息穷人。

这是一个巨大的怪圈，我们生活在信息爆炸的时代，而却成了信息穷人!

该怎样走出这个怪圈呢？

采用邹氏技术(邹晓辉先生发明的语义信息及真实文本的全域数码定位技术)，即可走出这个怪圈。

邹氏技术是一项以大量理论创新为前提的、把握住了当前知识信息处理发展方向的、具有很大应用前景和市场价值的专利技术。应该引起各方面的高度重视。

知识信息处理的关键是计算机理解语义信息，而这在目前仍然是世界性的难题。我国北京大学、北京邮电大学、北京语言文化大学、中科院的著名专家都对此进行了多方面的探讨，取得了不同程度的进展。

邹晓辉先生对语义信息理解进行了长达20余年的研究，提出人工智能、人类智能和协同智能关系的理论；义、文、物、意构成的知识分类体系；信息的基本定律［ 本真信息，唯一守恒；广义文本，对应转换；基因通式，序趣简美；特式特例，非非各平（即：非对称、非同步、各自平衡）］和文化基因通式等，形成了融智学新范式。我以为，融智学新范式是创立者长期研究、深思熟虑、融汇贯通的结果，是应该加以足够重视的。

邹晓辉先生不仅进行了理论创新，而且以应用为直接目的，形成了一种知识信息数据处理方法，以及核心产品 :1. 基于数学的通用操作系统和定制的文化基因文本数据库； 2. 为用户定制的操作应用系统和个性化的知识信息数据库。统称为邹氏技术。

邹氏技术的应用是多方面的，简略说来，有如下方面：

计算机领域，简化形式信息的处理，并形成统一的国际标准软件体系——全域数码定位体系，即：区别于现有的各种操作系统及其应用系统和知识信息数据库的融智新体系。

通信领域，简化各种具体知识信息传递方式，以默契通信的方式最大限度地提高通信效率及效能。

传感及传播领域，简化形式信息的输入和输出，消除人机界面知识信息数据交流的各种障碍。

人工智能领域，在专家系统和领域专家的工作基础之上，构成协同智能主体及网络，在全球网上实施知识系统工程和知识产权监管，自动测评网上交流的知识信息的价值。

新闻出版及广播电视领域，借助文化基因工程减少乃至消除冗杂文本，合理配置电子出版物与印刷出版物所占用的资源。

教学培训领域，借助人工智能和文化基因工程等方面的新技术及新设备，提高人力资源的开发效率及效能。

传统产业改造领域，对农业、工业、商业和服务业进行信息化改造，显著降低其生产、管理、

经营等方面的成本，使之更加精细化，进而，使其产品更容易适销对路。

目前，IT 业竞争激烈，IT 业发展初期仅靠先走一步而成功的可能性大大减少了，继之而起的将是核心竞争力的竞争。而企业最重要的核心竞争力是核心技术，邹氏技术即具有直接转化为产品的核心技术的特征。有远见的企业家、有志于在 IT 业发展者和各级决策者，应予以高度重视，否则，他人将先我而得之，悔则晚矣。
(1810)

系统科学专家（四川大学 陈雨思）有针对性的评语之三

系统科学之窗

系统科学谈天说地

主题： 邹晓辉的知识分类体系与知识系统的结构同一性

陈雨思 邹晓辉的**融智学新范式**提出**义、文、物、意**构成的**知识分类体系**。

核心会员

发贴数量: 323

来自:成都

注册日期: Dec 2000

邹晓辉指出：

义，指本真信息。如曲、棋、语言的机理（含：法则）。

文，指符号形象。如乐谱、棋谱、字或字母或动作等。

物，指载体载能。如琴、棋、传感器官（含使用过程）等。

意，指意识意向。如演奏者、下棋的人(作出的一系列选择)。

这个知识分类体系有什么意义呢？我们暂时把载体载能放在一边，这当然不是因为它不重要，恰恰相反，知识的载体是很重要的，如没有计算机这个载体，我们就无法处理信息。但是，现在我们要说知识本身，而不说知识的载体，所以把它放在一边。同时我们只以象棋为例子，因为我们中间喜欢下象棋的人肯定不少。

以象棋为例子，那么"义"就是指象棋的棋理。

象棋的棋理其实是很简单的，象棋棋子只有 32 个，棋盘上只有 90 个交叉点，象棋子就摆在这些交叉点上。象棋走法也很简单，就是人们常说的：车行直路象行田，马行斜日炮翻山，卒子过河横竖走，士象不离老王边。这些棋理，每

个人只要肯花半个钟头时间，就能记住。

虽然象棋的棋理很简单，但它是每一个下棋的人都必须遵守的，无论棋局怎样千变万化，这些棋理是不变的、守恒的。

以象棋为例子，那么"文"就是指象棋的棋谱。

已经成书的象棋谱究竟有多少，恐怕难以统计清楚。不过中国著名象棋谱倒是可以列出一个清单。

在古代，著名的象棋谱有《事林广记》、《百变象棋谱》、《桔中秘》、《梅花谱》、《竹香斋象戏谱》、《韬略元机》、《心武残篇》、《百局象棋谱》等。

在近代，著名的象棋谱有谢侠逊的《象棋谱大全》，内容包括《适情雅趣》、《烂柯神机》、《桔中秘》、《梅花谱》、《象局汇存》、《象局集锦》、《竹香斋象戏谱》、《弈乘》、《弈话》、《万国象棋》等棋谱和棋话。

在现代，著名的象棋谱有杨官龄的《中国象棋谱》，此书总结了近百年来中国象棋的理论和实践。

由此可见，如果与象棋的棋理比较，象棋谱涉及的内容要复杂得多，而且对于下棋的人来说，他并不是非要遵守这些象棋谱中的每一种下法不可，而是可以选择的。虽然有所选择，象棋谱揭示的一些下象棋的共同规律，人们还是要遵守的。

以象棋为例子，那么"意"就是指下象棋的人(作出的一系列选择)。

古往今来，下象棋的人究竟有(作出过)多少(选择)，恐怕谁也说不清楚(但是大家所作出的一系列选择的类型却可以概括出来)。象棋在中国有着悠久的历史。在 2000 多年前成书的《楚辞·招魂》中就有关于象棋的记载。据传说：楚汉相争时，韩信带兵攻打赵、齐等国，一段时间打仗，一段时间休整，在休整时作象棋以教士兵。现在象棋棋盘里的河界，还叫做"楚河汉界"。

在 2000 多年的象棋战争中，人类进行过多少次象棋对局，这恐怕是一个天文数字。这些都是人们的意识意向的体现，人们只要不违背象棋的棋理，就可以任意对局，而不一定要考虑下象棋的规律。

象棋这个例子给我们什么启示呢？

任何一个棋谱，总是由一些对局构成的，这些对局当然是从许多对局中选择出来的，选择的理由是进入棋谱的对局要有代表性，要反映各种对局的共性。所以说，棋谱是各种对局共性的体现。

在象棋对局中还有一个所有棋谱和对局都必须具有的共性，这个共性就是所有棋谱和对局都必须遵守棋理。所以说，棋理是所有棋谱和对局共性的体现。

说到共性，就使我们想起同态学中的同一性定义：同一性是指系统要素间的共同性质。那么，对于象棋对局系统来讲，它的棋理、棋谱和对局就都有一个同一性，而且有：

对局的同一性≤棋谱的同一性≤棋理的同一性

按照珠子模型，可以用（0，1）之间的一个数来表示系统的同一性。

因为棋理是所有棋谱和对局都必须遵守的，所以棋理的同一性是 1。

棋谱的同一性确定比较麻烦，这要看棋谱的代表性，如果棋谱的代表性比较强，它反映各种对局的共性就比较多，它的同一性就比较大，例如它的同一性可以是 0.6。如果棋谱的代表性比较弱，它反映各种对局的共性就比较少，它的同一性就比较小，例如它的同一性可以是 0.3。

对局的同一性确定更麻烦，这也要看对局的代表性，如果对局能够进入棋谱，它的代表性就比较强，例如它的同一性可以达到 0.6。如果对局的代表性比较弱，它反映各种对局的共性就比较少，它的同一性就比较小，例如它的同一性可以趋近于 0。

应用珠子模型，可以对于象棋对局系统的棋理、棋谱和对局进行精确的同一性分析，这种分析可以使我们对于象棋对局系统的结构有精确的了解，从而为处理象棋对局系统（例如用计算机下象棋）提供依据。

当然，下象棋只是我们举的一个例子，我们要讨论的是知识分类体系问题。从上面的讨论可以看到，邹晓辉的义、文、物、意知识分类体系是有它的科学依据的，这个科学依据就是知识系统的同一性。

义是知识系统中同一性最大的子系统，它的同一性是 1。在知识系统的变换过程中，它是不变的、守恒的，所以邹晓辉强调指出，**本真信息，唯一守恒。**

文是知识系统中同一性居中的子系统，它的同一性是（0，1）之间的一个数。在知识系统的变换过程中，它是可以变化的，不过，这种变换过程具有一定的代表性，可以在一定范围内通用。所以邹晓辉说，广义文本，对应转换；基因通式，序趣简美。

意是知识系统中同一性比较小的子系统，它的同一性也是（0，1）之间的一个数，但是趋近于 0。在知识系统的变换过程中，它是变动不居的，这种变换过程可能具有一定的代表性，也可能不具有代表性。所以邹晓辉说，特式特例，非非各平［即：非对称、非同步、各自平衡］）。

从知识信息的计算机处理出发，按照义、文、物、意来进行知识体系分类，是不错的。不过，对于人类的整个知识系统，我们还可以进行更广泛的同一性分析，以便能够对它的结构有更确切的了解。我曾经写过一篇文章，叫做《朱夫子的精神漂泊与参照系固执》，这篇文章其实就是对祖国主要的文化流派儒、释、道三家进行了简单的同一性分析，这对帮助我们理解祖国传统文化，应用计算机来研究祖国传统文化，以便古为今用，其意义是不言而喻的。

（注：同态学的内容集中在 110 万字《同态学文集》中。《同态学文集》目前只以光盘形式提供，联系邮箱：xiao_he_ping@263.net）

06-14-2003 20:33　注：其中褐色+括号的部分是邹晓辉批注的观点

系统科学专家（四川大学 陈雨思）有针对性的评语之四

作者	主题： 钱学森的研讨厅体系
chenyusi 特级会员 发贴数量：100 来自:成都 注册日期：Jun 2003	20 世纪 80 年代末，钱学森提出了处理开放的复杂巨系统的方法论，这就是"从定性到定量的**综合集成方法**"。在 90 年代，钱学森又进一步提出"从定性到定量**综合集成的研讨厅体系**"。 研讨厅体系的实质是专家体系、数据和信息体系以及计算机体系三者的有机结合，构成一个高度智能化的人机结合系统。 从思维科学角度来看，人脑和计算机都能有效处理信息，但两者有极大差别。人脑思维包含逻辑思维和形象思维。而创造思维则是逻辑思维和形象思维的结合。今天的计算机在逻辑思维方面，确实能做很多事情，甚至比人脑做得还好，已有很多科学成就证明了这一点，如著名数学家**吴文俊**先生的定理机器证明。但在形象思维方面，现在的计算机还不能给我们比较好的帮助，至于创造思维就只能依靠人脑了。既然机器(计算机)有机器的优势，人脑有人脑的优势，把两者结合起来就更有优势。所以人机结合以人为主，优势互补、相辅相成，人帮机、机帮人和谐地工作在一起。机器能做的尽量由机器去完成，极大地扩展人脑逻辑思维处理信息的能力。通过人机结合以人为主进行信息与知识的综合集成，这里包括不同学科、不同领域的科学理论和经验知识、定性和定量知识、理性和感性知识，通过人机交互、反复对比、逐次逼近，达到从定性到定量的认识。这个方法的成功应用，就在于发挥这个系统的整体优势、综合优势和智能优势。这个人机结合系统在思维能力和创造性方面，比起单纯靠人(专家)或机器都有更强的优势。它能把人的思维、思维的成果、人的经验、知识、智慧以及各种情报、数据和信息统统集成起来，从多方面的定性认识上升到定量认识。 综合集成方法和研讨厅体系作为科学方法论，它的理论基础是思维科学，方法基础是系统科学与数学，技术基础是以计算机为核

心的现代信息技术，哲学基础是马克思主义实践论和认识论。

信息的综合集成可以获得知识，信息和知识的综合集成可以获得智慧。从这个意义上说，综合集成方法和研讨厅体系是人机结合的知识生产体系，是知识生产力和精神生产力，它使人们由过去完全依靠人脑进行知识生产转变为人脑、电脑相结合的知识生产方式。这是当前这场信息革命对人类社会发展影响的一个重要方面。

03-12-2004 17:45

chenyusi

特级会员

发贴数量：100

来自:成都

注册日期：Jun 2003

邹晓辉的**协同智能**似乎**与**研讨厅体系**有内在的一致性.**

03-12-2004 17:59

语言及语义信息的统一参照系

摘要

提出义、文、物、意的融智概念体系，是对传统的意义理论以及人类知识概念体系进行彻底变革。融智概念体系以及相应的数学模型是语言以及语义信息的统一参照系的最佳形式。

引言

为什么这样说“提出义、文、物、意的融智概念体系是对传统的意义理论以及人类知识概念体系进行的彻底变革，融智概念体系以及相应的数学模型是语言以及语义信息的统一参照系的最佳形式”呢？本文通过回答以下问题来阐述其理由。

什么是融智概念体系？它涉及那些基本概念、推理法则和层次结构？能举例说明吗？（内容）

传统的意义理论以及人类知识概念体系，是一个什么状况？为什么说语言以及语义信息的交流必须要有一个客观的统一参照系？（必要性和重要性）

融智概念体系以及相应的数学模型所表达或反映的统一参照系，为什么是客观的？还能够找到比它更简单且完善的客观统一参照系吗？（必然性和可行性）

一

所谓融智概念体系，是由义、文、物、意构成的知识分类体系。在此，义，指本真信息；文，指

符号形象；物，指载体载能；意，指意识意向。其中，文、物、意，统统被视为展示本真信息的广义文本。与现有的知识分类体系相比，其本质特征在于对义与意进行严格区分。由此可见，义、文、物、意是构成该体系的基本概念。它们的含义可通过以下事例做进一步的通俗的说明，即：

原理，如杯子的机理，是本真信息，属于义的范畴；展示其机理的文化形式，如杯子的图纸，属于文的范畴；展示其机理的物化形式，如具体的杯子，属于物的范畴；智能主体的选择，如杯子的构造及外观的设计构想，属于意的范畴。

融智概念体系与其他概念体系新、旧区别的关键在于：

旧范式，对义与意，不作区分。例如：现有的语义学、意义理论和语义信息理论，不仅都没有明确地区分义与意，而且，总是用意义的概念把义与意混为一谈。值得注意的是，不用汉语的字本位的观点和方法，是不足以凸显这个问题的。

新范式，对义与意，严加区分。例如：《一种知识信息数据处理方法及产品》的原理，不仅对义与意作了严格的区分，而且对义、文、物、意作了严格而又明确的定义。

如果说，文与物，涉及图、文、数、表、音、像、立体、活体等形式信息，那么，义与意，则涉及义、意、意义等语义信息。众所周知，个人计算机革命和通信革命的成果主要体现在形式信息处理方面，其困难则主要发生在语义信息处理方面。

试问：旧范式在语义概念本身都存在问题的情况下，如何解决语义信息领域面临的难题？又如何解决语义信息和形式信息交错的难题？何况形式信息领域本身面临的难题并没有彻底解决。

新范式则把义、文、物、意四个分类系列作为协同智能主体进行定性分析的分类基础，区别于现有的旧范式及其分类体系。

二

下面通过汉语与英语的具体对比，举例说明融智概念体系的基本概念、推理法则和层次结构。

1、从“义”的范畴看，汉语与英语各自的机理是不完全相同的。如果说普通语言学着重研究的是各种语言的共性，那么，汉语语言学着重研究的就应该是汉语这种语言的个性，而计算语言学的主攻方向就应该是计算机如何自动处理自然语言——包括自动化标注或识别、定位或理解、再现或表达。

2、从“文”的范畴看，汉语的字和英语的词，是截然不同的符号形式体系（这是众所周知的）。

3、从“物”的范畴看，汉语的字和英语的词的发音状态及过程，以及书写形式，也都是截然不同的物态形式体系（这也是众所周知且显而易见的）。

4、从“意”的范畴看，汉语的字和英语的词，虽然都可以指谓相应的“物”乃至“义、文、物、意”，但长期使用汉语的字或英语的词的人或民族必然会产生截然不同的识别、理解和表达的习惯（这更是显而易见的）。

5、从“义、文、物、意”诸范畴看，汉语的字和英语的词的最大的不同是：在“基本笔画、偏旁部首、字、字组、句、段、篇章”中，以“字”为界，字以前是非线性结构而且是单音节，字以后是线性组合而且是多音节；在“字母、词素、词、词组、句、段、篇章”中，以“词”为界，词前后都是线性组合，而且词本身既可以是单音节也可以是多音节。由此而产生其它一系列不同，涉及具体语言文字的机理法则，体现为：符号形式、发音习惯和思维表达习惯等具体的特征部分。

其中的推理机制及层次结构，根据信息基本定律和文化基因通式，可作如下表述：

基本笔画和字母是子全域的基因文本元素；

偏旁部首、字、字组（或词素、词、词组）以及随后的句、段、篇章等，都属于超子域的基因文本元素组合。

其中，汉语的字和英语的词最适合分别被视为各自语言文字体系中具有语义特征的基本结构单位，因为，在它们之后的字组和词组、句、段、篇章等语言单位都是由它们的线性组合形式所构成的，只要代码化就可被计算机自动处理。

英语的计算机自动处理，国外的计算语言学理论、软件工程和知识工程等领域有大量成果可供借鉴，在此就不一一列举了。

但是，从中文信息处理的角度看，汉语的字本位是顺理成章的。以外来的词或词组作为汉语的基本结构单位是人为强加的，不符合汉语的机理和法则。至于采用字组（词和词组）、句子（包括小句或单句和复句）作为汉语的基本结构单位，也都不合逻辑。换句话说，汉语的基本结构单位只能是字。其它的所谓本位说充其量只能是一种临时的过度策略或战术应急手段，绝不具有战略的地位。

至于汉语的字和英语的词之间，进一步包括它们之后的字组和词组、句、段、篇章等语言单位在内，如何进行比较或定位的问题，我认为都可以从子全域与超子域之间的演绎关系以及已知域与目标域之间的比对关系的计算机自动处理中找到答案。

三

传统的意义理论以及人类知识概念体系，是一个什么状况？以下对“语义三角”进行深入的分析和透彻的表述。

意义，作为语义信息的核心概念，由于它暗含：“意=义”这样一个假设，因此，造成了极大的混乱。我认为：这正是传统的意义理论以及人类知识概念体系中语义或语义信息等概念及其表述形式的根本问题之所在，这也是造成人类及人工智能主体在对语义信息进行定性及定量分析处理方面长期存在的一系列瓶颈问题始终得不到根本解决的原因之所在。

毋庸置疑，“意≠义”是常例，而“意=义”只是其特例。意与义，决不能被简单地混为一谈。这就是为什么有必要提出“义、文、物、意”的融智概念体系的最基本的理由，也是为什么有必要对现有的意义理论以及与之相应的人类知识概念体系进行彻底的变革的根本理由。在“义、文、物、意”的融智概念体系中，除了用义表示本真信息之外，文、物、意，都属于义的衍生形式或派生现象。其中，文，指符号形象；物，指载体载能；意，指意识意向。

从哲学的角度看，本体论的形成、认识论的发展、语言学的转向，哲学领域这三次大的进步,实质上是从物、意、文三个方面分别形成了相应的有关其内在含义的立论体系。虽然它们都试图向本真信息迈进，但是，由于历史的因素或时代的局限等种种原因，至今，始终没有形成完整的义或本真信息的立论体系。

从科学的角度看，科学的各门学科，虽然都是以探求本质、机理及法则等本真信息为己任，但是，至今仍不具备形成完整的义或本真信息的立论体系的条件。因为，科学旧范式的特征使得各个学科“只见树木而不见森林”。到目前为止,各门学科实际上都是在探索现象背后的本质，即：支配物、意、文的义。但是，非常遗憾，其结果往往是支离破碎难以整合。

至于艺术、技术、工程等领域，虽然都有各自相应的思维表达形式，但是，却在根本上受到哲学和科学的思维表达方式的影响。

表达式，涉及广义真实文本，包括:数、文［狭义的文本（文字）］、图、表、音、像、立体、活体（在此，物，可以被视为特殊的广义的文本）。

智能主体（包括自然人、机器人、协同智能主体等）指谓的对象（包括形式及其内容），可以是义、文、物、意的任何一个方面的具体内容或形式。

从通俗的例子来看：曲、棋、语言的本质或机理或法则等，可视为义；乐谱、棋谱、文字等，可视为文；:琴、棋、器官或装置等，是物；:演奏者、下棋的人、智能主体本身的选择，可视为意。其中，只有义是唯一守恒的，而文、物、意都是可变的。

如何判断并选择变化的形式或了解与认识变化的法则？直接关系到能否有的放矢地调整智能主体自身的行为方式（包括说、写、做），同时，相应地改造主、客体之间的关系(包括：主体与主体、客体与客体、主体与客体之间的相互关系)。

由于智能主体（包括自然人、机器人、协同智能主体等），是特殊的整合物，因此，对知识信息的识别、理解、表达，实际上就是智能主体上、下载知识信息或进行数据结构转换的过程。既然如此，转载信息的基本单位是汉字、英词或其它什么形式，这并不是问题的关键之所在。

或许对自然人而言，分别以字或词为基础，能构成不同的思维方式及相应的句法形式和表达体系。但是，对机器人或系统而言，关键是如何对构成字或词的基本元素的识别和计量的问题，因为各种类型

和层次的基因文本元素组合，都可以通过相应的算法或排列组合进行处理或理解与表达。

四

融智概念体系以及相应的数学模型所表达或反映的统一参照系，为什么是客观的？还能够找到比它更简单且完善的客观统一参照系吗？

众所周知，学术界（其中数学是唯一的例外）这种“公说公有理，婆说婆有理”的现象是经常发生的事。对此，人们的态度通常是见怪不怪。

为什么人们见怪不怪呢？为什么数学会是唯一的例外呢？又有多少人深思过这类问题的答案呢？我认为：这几乎是一个无法说明白的问题，除非能够跳出现有理论的框框。

这些“公说公有理，婆说婆有理”的人，都没有真正摆脱现有理论框架的约束。因此，只好把矛盾问题搁置在那里。学术界的许多根本性的问题也都是这样被耽搁的。

如果两个智能主体之间在传递知识信息数据的过程中，各自依据的参照系不同，即：各自进入了“编号排位”不同的“分剧场”就算双方的“号”与“位”的形式都是一样的，但是，实际上却根本不搭界。现在，包括过去和近期的未来，各个智能主体之间的知识信息交流或数据交换，往往就属于这种“不搭界”的状态。

如果各个智能主体之间，在进行知识信息交流时，能够依据统一的参照系，那么，交流和理解的许多困难都将迎刃而解。人类的确是在向这个方向努力的。但是，由于作为共同参照系的科学范式以及相应的语言体系本身（特别是人类的概念体系）处于变化之中，加之海量信息和知识爆炸，使得人们往往只有招架之功而没有还手之力。

由于近、现代以来形成的相对时空观以及相应的逻辑体系，使得相应的思维体系、语言体系和软件体系都充分表现出各自为政的多样化冗余特性。因此，当“序、趣、简、美”且非常适用的统一参照系展现在面前之时，人们也往往会视而不见，即：对它的存在不理解，也感觉不到。

各个国家、各个产业、各个行业、各个企业、各个人、各个学科，似乎都在忙于建立或维护各自的相对参照系，并在这种各自为政的忙碌之中随着生命周期的变换而消逝。

义的形而上的特性、文的形而中的特性、物的形而下的特性，都受到意的无形或变化不定的特性的影响或干扰而难以为人们所认识和把握。

至今为止,还没有其他任何人提出过这种具体的可操作的统一参照系。不知牛顿后半生追寻的“神”和爱因斯坦晚年试图证明的“统一场”是否就是上述这种统一参照系?

我认为：融智概念体系以及相应的多元数系就是上述这种统一参照系的最佳理论形式。

由此构成的文化基因通式[代数表达式（a+bi&…）几何表达式(t&…,x,y,z)]，既是绝对时空观和相对时空观统一的最简形式，又是唯一能够集中表达其它所有文化形式的基因元素及其组合的万能公式。

例如：全域数码定位系统[代数表达式（a+bi&…）与几何表达式(t&…,x,y,z) 的物化形式，或：义、文、物、意四大范畴结合的典型实施例]，就是上述这种统一参照系的工程化形式，它对文化基因元素及其组合的计量或“编号排位”就是协同智能主体对知识信息的表达，反过来的“对号入位”则是对知识信息的识别和理解。

五

关于字本位的一点补充说明。

下面直接应用《融智学（新范式）》义、文、物、意四个基本范畴，对汉语和英语进行比较分析，从而说明采用“字”作为汉语的基本结构单位的合理性。

字的非线性结构与词的线性结构之间的区别，是汉语与英语（等拼音文字语言）之间最根本的区别，其它区别几乎都因此而产生，例如：思维方式（包括联想和推理的具体形式，例如《易经》的“两点论”与亚里士多德的“三段论”）及其表达方式（包括语法的具体形式，例如汉语的“语义句法”与英语的“语形句法”）的区别。

汉语的字和英语的词，都属于“文”的范畴；两者都是用于表达“意” 或称谓“物”的语言文字

工具。至于智能主体的“意”是否与客观的“义”吻合，或究竟称谓的是“物”、“文”、“义”还是“意”？这已经超出了现有的语言学的范围。

因此，不发现并发明一种区别于现有的哲学以及科学旧范式的融智学新范式，是不可能高屋建瓴地把整个问题的本质搞清楚的。

因为，仅仅在汉语的字和英语的词所属的“文”的范畴之内，来评价各种观点，最好的、最公平的、最有益的结果，只能是各抒己见。如果要真正地推进人类的认识就必须突破现有的哲学以及科学（包括语言学和计算语言学）旧范式的理论框架。

注:至于融智学理论以及文化基因公式如何具体地描述或表达,特别是如何利用该统一参照系为人类做事的具体方法和形式,在《一种知识信息数据处理方法及产品》(专利已提前公开)的发明专利说明书中有详细说明及实施例。

参考数据

1、北京大学中文系 徐通锵《语言论》1997年10月东北师范大学出版社

2、邹晓辉《一种知识信息数据处理方法及产品》CN1274895A国家知识产权出版社2000年

3、邹晓辉《融智学（新范式）》系统科学之窗论文专区：

附录：

一、语义三角，即：semantic triangle：concept(thought)，symbol(word)，referent(thing).。《the meaning of meaning》by C.K.Ogden and I.A.Richards 1923年

二、中国智网首页（ ） 文化基因工程（语义信息处理）公开发表时间2000-12一篇重要文章：《语言以及语义信息的统一参照系》原创者兼著作权所有人：邹晓辉(zouxiaohui）；熵.信息.复杂性网首页（ ）2001年6月转载。

三、这次修订，重点参考了以下数据：

1、邹晓辉《从融智学的观点看汉语的一个基本理论问题——与徐通锵和陆俭明两位教授商榷》《现代汉语》Modern Chinese Board 2001年9月

2、陈雨思《克服不确定,发展系统科学》、《信息自组织与知识信息处理——与邹晓辉交流》系统科学之窗论文专区

3、易绵竹“计算语言学探索（系列文章）”《位语法理论与应用》黑龙江人民出版社1999年

4、周斌武、张国梁《语言与现代逻辑》复旦大学出版社1996年12月第一版

5、陆俭明“汉语语法研究所面临的挑战”《计算语言学文集》（余士文等）2000年12月 北京大学计算语言所

6、邹晓辉《关于“克服不确定,发展系统科学”与陈雨思的交流》系统科学之窗论文专区

语言学和计算语言学专家（北京大学徐通锵）有针对性的评语之一

专家评语：

“义、文、物、意概念的提出，我觉得是有价值的，但如何阐述？抓住什么样的核心？需要深入推敲。这四个概念的关系，据我的理解，‘义’应是客观存在的事物的结构机理，或者说是客观规律，其运转规律不以人的主观意志为转移；‘意’是主观对“义”的认识或理解，‘文’与‘物’只是这种认识外化的表现形式。区分‘义’与‘意’是很必要的，现代语言学也已意识到这种区分的必要，其具体的表现形式就是注重功能的研究。如何将这种区分进行理论上的阐述，还需学界的努力。”

北京大学中文系基础语言学研究室(原)主任**徐通锵**教授

（全国普通高等学校人文社会科学研究十五规划纲要 语言学 咨询组负责人）

原文附件：

晓辉先生：

几次来函和大著都已先后收到，谢谢。先生提出了一些重大的问题，鄙意是目标过大，恐怕难以实现预期的目的。因为我们研究的领域不同，很多东西我不懂，只能供先生参考。

义、文、物、意概念的提出，我觉得是有价值的，但如何阐述？抓住什么样的核心？需要深入推敲。这四个概念的关系，据我的理解，“义”应是客观存在的事物的结构机理，或者说是客观规律，其运转规律不以人的主观意志为转移；“意”是主观对“义”的认识或理解，“文”与“物”只是这种认识外化的表现形式。区分“义”与“意”是很必要的，现代语言学也已意识到这种区分的必要，其具体的表现形式就是注重功能的研究。如何将这种区分进行理论上的阐述，还需学界的努力。

先生采纳字本位理论，我自然很高兴。从几次来文看，觉得字在解决融智体系中相关问题的地位和作用，说得不清楚。您曾跟我说过，您看到有关字的文章后，原来思想上一些不清楚的问题，相接通电路一样，一下子就清楚了。我很看重您的这种感觉。如果字在计算机运算中能发挥它的潜能，在融智体系中能解决现行的语法理论不能解决或解决不好的问题，那对科学的发展就很有意义。看了先生的几次来函和稿件，只看到字的重要性，主张放弃词，以字为本位，但我好象始终没有看到它在解决融智体系的问题的作用。可能这方面涉及很多具体的问题，不是先生的信函所能解决的，这里只是提出来供参考。评判一种理论的优劣，主要是根据简明、解释力、可操作性三个标准，希望先生能在这方面作出新的努力，结合科学的发展和融智体系的目标，实事求是地进行一些评述和阐释。

另外，有一个小问题附带说一下，就是在9月5日来函的附件中说，由于词的引入，汉语中产生和增加了大量的字组。这一说法不确切。字组的产生和发展是汉语自身自我调整的结果，与词的引入无关。

先生为了照顾我阅读我在您的网上看不到的数据，特地发来的一个很大的附件，很遗憾，据瑞星防毒软件提示，有病毒，我只能删除。我的自然科学水平低，即使打开了附件，也很可能看不懂。这就算了吧。我们以后可以多讨论一些基础理论问题。

以上意见仅供参考。

徐通锵

2001， 9，11

徐先生：

您这次来函明确指出：

我的文章“提出了一些重大的问题”，同时，又认为：我（或我们？）的“目标过大，恐怕难以实现预期的目的。”并且，谦虚地说：“因为我们研究的领域不同，很多东西我不懂，只能供先生参考。”

对此，我的意见是这样：

既然“提出了一些重大的问题”，就不要怕“目标过大”，更加不要怕“难以实现预期的目的”，因为，对探求真理的人们来说，各个人的力量（包括时间、精力或生命）都是非常有限的，但是，绝不能因此而畏惧困难，反而应该有愚公移山的精神。正如您在《语言论》中所说的那样——即使失败也可给后人提供警示。因此，我认为：这里有两个问题（涉及目标与方法两方面）值得进一步探讨，即：

1、您认为我“提出了（哪）一些重大的问题”？为什么您认为它们是“重大的问题”？

2、您说：“目标过大”，是否主要针对个人的力量（包括时间、精力或生命）非常有限而言？或者说：您有使目标更加具体化，或您有使该目标分解得更加容易实现的具体办法？

接着，您说：

“义、文、物、意概念的提出，我觉得是有价值的，但如何阐述？抓住什么样的核心？需要深入推敲。这四个概念的关系，据我的理解，义应是客观存在的事物的结构机理，或者说是客观规律，其

运转规律不以人的主观意志为转移；意是主观对义的认识或理解，文与物只是这种认识外化的表现形式。区分义与意是很必要的，现代语言学也已意识到这种区分的必要，其具体的表现形式就是注重功能的研究。如何将这种区分进行理论上的阐述，还需学界的努力。”

对此，我的意见是这样：

既然“有价值”，就应该尽快向其他有关专家（包括您的学生及助手）和领导说明，以便新思想、新观点、新理论和新方法的推广普及。至于“如何阐述？抓住什么样的核心？需要深入推敲”是指我的理论文章（是《融智学（新范式）》还是该理论的应用文章如：《语言及语义信息的统一参照系——给一本中国计算语言学文集的修订稿》）还是您的考虑？

对义、文、物、意这四个概念的关系，您的理解——“义”应是客观存在的事物的结构机理，或者说是客观规律，其运转规律不以人的主观意志为转移；“意”是主观对“义”的认识或理解——是对的，但是，“文”与“物”不只是这种认识外化的表现形式，因为，即使它们不被人们所认识也存在。

区分“义”与“意”是很必要的，但是，这种区分的必要性绝不限于某一个学科领域（包括语言学）。现代语言学可能意识到了这种区分的必要，但是，如果说“其具体的表现形式就是注重功能的研究”，那么，我认为其认识的深度和广度还远远不够。更不用说“将这种区分进行理论上的阐述”了，的确“还需学界的努力”。

值得强调的是，您说：

“先生采纳字本位理论，我自然很高兴。从几次来文看，觉得字在解决融智体系中相关问题的地位和作用，说得不清楚。您曾跟我说过，您看到有关字的文章后，原来思想上一些不清楚的问题，像接通电路一样，一下子就清楚了。我很看重您的这种感觉。如果字在计算机运算中能发挥它的潜能，在融智体系中能解决现行的语法理论不能解决或解决不好的问题，那对科学的发展就很有意义。看了先生的几次来函和稿件，只看到字的重要性，主张放弃词，以字为本位，但我好象始终没有看到它在解决融智体系的问题的作用。可能这方面涉及很多具体的问题，不是先生的信函所能解决的，这里只是提出来供参考。评判一种理论的优劣，主要是根据简明、解释力、可操作性三个标准，希望先生能在这方面作出新的努力，结合科学的发展和融智体系的目标，实事求是地进行一些评述和阐释。”

对此，我的意见是这样：

“字在解决融智体系中相关问题的地位和作用”，主要体现在：

1、汉语的字，位于非线性结构与线性结构之间的基本语言单位，可以使融智学新范式的四个基本概念表达的非常简单、明了；

2、例如，对意与义的区分，特别是对它们与意义（meaning）的区分，比采用由字的各种组合而派生的字组和句子更加精练；

3、汉语的字，作为一种表意与表义的自然语言，必然在解决语义信息的问题或“语义泥潭”方面作出特殊的开创性贡献。

我曾跟您说的那种感觉，就像您的名字那样铿锵有力——因为您“有关字的文章”把汉语的精华表达得淋漓尽致。我认为：不仅是汉语理论的“一些不清楚的问题”“一下子就清楚了”，而且还对寻找汉民族在未来发展的目标方向或国际定位方面有所贡献，这正是《融智学（新范式）》及其文化基因工程或协同智能主体的概念体系寻求的语言表达或先期推广的独特优势。

既然您已经看到了我“几次来函和稿件，只看到字的重要性，”那么，为什么不谈一谈字的定义问题呢？别忘了陆俭明教授对此的公开质疑是有一定依据的。何况字的定义的确是您的理论能否立得住的一个重要基础。

有必要指出：现行的语法、语义和语用等旧的学科理论范式，存在根本性和深层次的问题，这在形式化程度较高的语言（如英语）中的隐蔽性比在形式化程度较低的语言（如汉语）中更加明显。现代计算机采用的基因文本元素由ASCⅡ体现得非常充分，因此，后续派生的词、词组或短语、句子、段落、篇章等基因文本组合，都容易为算法及程序加以表达或转述。但是，对汉语而言，在形式化、结

构化和程序化方面，不仅没有这种优势，而且，还存在中文信息计算机处理的瓶颈（不采用新范式是不可能从根本上得以突破）。由此可见，对中文信息的计算机处理必须同时找到汉语的自然语言学和计算语言学的理论及实践的突破口。我认为：您的《语言论》主要在自然语言学方面作了大胆而有益的探索，而我的融智学新范式则由于注重人机协同而与现行的自然语言学和计算语言学产生了交叉或部分重叠。

根据您说的“评判一种理论的优劣，主要是根据简明、解释力、可操作性三个标准，”同时，根据易绵竹教授指出的“按照20世纪的学术思想，任何一门科学都是一个知识体系，它需要有基本概念、推理法则和层次结构，凡是不能纳入这个知识系统的知识，就不能称之为科学知识。”我认为：《融智学（新范式）》和《一种知识信息数据处理方法及产品》等在系统科学之窗论文专区

或

融智网 或 上面

发表的一系列文章，基本上公开了符合上述要求的内容，只是在具体的表达形式方面与旧范式的各个学科之间的关系有待进一步详细说明。

最后，您说：

“另外，有一个小问题附带说一下，就是在9月5日来函的附件中说，由于词的引入，汉语中产生和增加了大量的字组。这一说法不确切。字组的产生和发展是汉语自身自我调整的结果，与词的引入无关。”

对此，我的意见是这样：

您的这个观点，我基本上表示同意。应该说：现代汉语中产生和增加了大量的灵活的字组——采用字组的方式不限于古代汉语中的成语、谚语、俗语，如：大量的短语、术语、习语等等。“字组的产生和发展，”的确主要“是汉语自身自我调整的结果”，同时，也“与词的引入”有关，如：与大量的外来词对应的字组的形成或采用。

以上讨论一些基础理论问题，希望多听您的意见！

邹晓辉 2001-9-12

语言学和计算语言学专家（洛阳外国语学院易绵竹）有针对性的评语之二

专家评语：

“融智学新范式提炼出协同智能主体的概念体系具有原创性，想必对自然语言语义信息的处理将引发一场革命。”

中国人民解放军洛阳外国语学院计算语言学研究室主任易绵竹教授（留学回国博士）

原文附件：

邹先生：您好！

感谢发来大作及与徐通锵先生的通信讨论文章，您的《语言及语义信息的统一参照系》一文**将收入来年出版的**计算语言学文集。

有关义项分类体系和计量方法，我目前考虑还欠周详，主要工作分派给几位研究生在做。我认为，对动词的义项分类以“状态”（to be）、“关系”（to be relative ）、“动作”（to do）作为原型特征，具有较大的概括性。而对名词进行义项分类应主要考虑以下的两对次范畴化特征：表人/表物；具体/抽象。至于计量方法问题，我以为语料统计的结果是可信的，这涉及技术性的问题，将交由软件开发人员负责。

当前语义研究的理论方法还需融合统一，您创立的融智学新范式提炼出协同智能主体的概念体系具有原创性，想必对自然语言语义信息的处理将引发一场革命。最近，我在探讨信息处理用概念词典、语义词典的工程实施问题，11月份参加中文信息学会理事会时，我可能就此问题参与

学术交流。您愿否赴会？会期及地点：11月11-13日，北京中苑饭店。联系方式：100080，北京8718信箱 中国中文信息学会

E-mail: cips@admin.iscas.ac.cn Tel: 010-6256-2916

顺便向您推荐一位具有哲学背景研究人员的论文(附件)，也许对您有用。

祝中秋国庆双节快乐！

易绵竹 2001-09-25

附录（以上修订稿依据的**原创文章**）：

说明：本文取自邹晓辉的融智网站，特表示感谢--编者2001，6，23

-------（“熵信息复杂性”首任主编 张学文）

语言及语义信息的统一参照系

摘要：提出义、文、物、意的融智概念体系,是对传统的意义理论以及人类知识概念体系进行的彻底变革。融智概念体系以及相应的数学模型是语言以及语义信息的统一参照系的最佳形式。

关键词：

语言,是智能主体之间进行知识信息交流的工具。文字则是语言的一种表达形式。现代汉语这种语言文字形式体系,基本上是自古代汉语的方块文字的延续和简化而来的,由于近代100多年来受到英文等拼音文字的影响,中西两类文化或文明的融合,形成了现代汉语的这种字词并行的格局。

古代汉语的基本语言单位是字。古代及现代的英语等西语的基本语言单位是词。现代汉语由于吸收了词的概念,许多古字则可以不用了,因为,无论是采用字还是采用字组做词,都可大大降低造字的需求数量，,即：组字成词或直接用字做词而不必再像古文那样造许多繁杂的字。这是引进英语等西语的表达形式的好处。

遗憾的是,字与词分庭抗礼的局面,事实上阻碍了现代汉语自身的理论体系的建立或完善。可以说,以字为基础建立的″语义句法″和以词为基础建立的″语形句法″{1}各有千秋,而且,它们各自都有相应的语言事实做支持。这就造成了现代汉语同时受两套思维模式、两种句法体系和两类表达形式体系所支配的现实格局。

不论是从有利于人类这种智能主体更好地理解现代汉语的角度,还是从有利于人工智能主体更准地识别、理解和再现或表达现代汉语的角度,都必须正视上述事实。

从中文信息处理的角度看,以字为本位则更加便于机器识别单个的字,而以词为本位却不便于机器识别或切分词,因为,机器难于区分独字词、二字词和多字词。从理解和表达的角度看,无论哪一种或哪一级的语言单位,都能从″语义三角″的各个方面进行分析和表述。

首先,意义,作为语义信息的核心,由于它暗含：“意＝义”这样一个假设，因此,造成了极大的混乱。我认为:这正是语义这一概念或语义信息的定性及定量问题始终得不到根本解决的原因之所在。

因为,″意≠义″是常例,而″意=义″只是特例。所以,意与义是决不能简单地混为一谈。为此,有必要提出义、文、物、意的融智概念体系,对传统的意义理论以及人类知识概念体系进行彻底的变革。在此,义,指本真信息:文、物、意,都属于义的衍生形式或派生现象,其中,文,指符号形象:物,指载体载能:意,指意识意向。从哲学的角度看,本体论的形成、认识论的发展、语言学的转向,这三次大的进步,实质上是从物、意、文三个方面分别形成了相应的立论体系。虽然它们都试图向本真信息迈进,但是,由于历史或时代的局限性和种种原因,至今,始终没有形成完整的义的立论体系。

例如：科学,虽然是以探求本质、机理及法则等本真信息为己任,但是,至今仍不具备形成完整的义的立论体系的条件,因为，科学的特征使得各个学科“只见树木而不见森林”。到目前为止,各门学科实际上都是在探索现象背后的本质,即：支配物、意、文的义,非常遗憾的是：其结果往往支离破碎而

难以整合,因为,哲学、科学、艺术、技术、工程,都有各自相应的表达形式。

其次,这里所说的表达式,涉及广义真实文本,包括:数、文［注:狭义的文本（如：文字）］、图、表、音、像、立体、活体（注：在此,物,可以被视为特殊的广义的文本）。

在此,指称的对象,可以是义、文、物、意的任何一个方面的具体内容与形式。从通俗的例子来看：曲、棋、语言的本质或机理或法则等,可视为义；乐谱、棋谱、文字等,可视为文；:琴、棋、器官或装置等,是物；:演奏者、下棋的人、智能主体的选择,可视为意。其中,只有义是唯一守恒的,而文、物、意都是可变的。

只有判断并选择变化的形式,了解与认识变化的法则,才能有的放矢地调整智能主体自身的行为方式,同时,相应地改造主、客体之间的关系(包括：主体与主体、客体与客体、主体与客体之间的相互关系〉。

由于智能主体,包括自然人和机器人或系统,是特殊的整合物,因此,对知识信息的识别、理解、表达,实际上就是一个智能主体上、下载知识信息的过程。既然如此,转载信息的基本单位是字或词并不是问题的关键。

或许对自然人而言,分别以字或词为基础,能构成不同的思维形式及相应的句法形式和表达形式体系。但是,对机器人或系统而言,关键是如何对构成字或词的基本元素的识别和计量的问题。

全域数码定位系统对文化基因元素及其组合的计量或"编号排位"就是智能主体对知识信息的表达,反过来的"对号入位"则是对知识信息的识别和理解。

如果两个智能主体之间在传递知识信息数据的过程中,各自依据的参照系不同,即：各自进入了"编号排位"不同的"分剧场",就算双方的"号"与"位"的形式都是一样的,但是,实际上却各是各的一一根本不搭界。现在,包括过去和近期的未来,各个智能主钵之间的知识信息交流或数据交换,往往就是属于这种"不搭界"的状态。

如果各个智能主体之间,在进行知识信息交流时,能够依据统一的参照系,那么,交流和理解的许多困难都将迎忍而解。人类的确是在向这个方向努力的。但是,由于作为共同参照系的科学范式以及相应的语言体系本身（注：特别是人类的概念体系）处于变化之中,加之海量信息和知识爆炸,使得人们往往只有招架之功而没有还手之力。

由于近、现代以来形成的相对时空观以及相应的逻辑体系,使得相应的思维体系、语言体系和软件体系都充分表现出各自为政的多样化冗余特性。因此,即使序、趣、简、美且非常适用的统一参照系展现在人们面前,人们也往往会视而不见。即：对它的存在不理解,也感觉不到。

各个国家、各个产业、各个行业、各个企业、各个人、各个学科,似乎都在忙于建立或维护各自的相对参照系,并在这种各自为政的忙碌之中随着生命周期的变换而消逝着。

义的形而上特性、文的形而中特性、物的形而下特性,都受到意的无形或变化不定的特性的影响或干扰而难以为人们所认识和把握。

至今为止,还没有其他任何人提出过这种具体的可操作的统一参照系。不知牛顿后半身追寻的"神"和爱因斯坦晚年试图证明的"统一场"是否就是上述这种统一参照系?

我认为：融智概念体系以及相应的多元数系就是上述这种统一参照系的最佳理论形式。

由此构成的文化基因通式就是绝对时空观和相对时空观的最简统一形式,又是唯一能够集中表达其它所有文化形式的基因元素及其组合的万能公式。

参考文献

1、北京大学徐通锵《语言论》1997年10月东北师范大学出版社

2、邹晓辉《一种知识信息数据处理方法及产品》CN1274895A国家知识产权出版社

注:至于该理论以及公式如何具体地描述或表达,特别是如何利用该统一参照系为人类做事的具体方法和形式,在《一种知识信息数据处理方法及产品》(专利已提前公开)的发明专利说明书[2]中有详细说明及实施例。

协同智能计算语言数据库的设计方法

技术领域

本发明属于语言信息处理技术领域，进一步是协同智能计算语言数据库的设计方法。

背景技术

2000年5月31日申报的“一种知识信息数据处理方法与产品（发明专利申请号001093800公开号1274895A）”和稍微晚些时候发表的“融智学（新范式）”（系统科学之窗论文专区），虽然定义并列举了文化基因的子全域与超子域及其进化阶梯的各个层次形式（以下简称：进阶层式），但是，却没有具体展示并详细分析基因文本元素及其组合形式，例如，没有说明汉语与英语在这方面是怎样区分的。

后来，我发现北京大学中文系教授徐通锵先生提出的“字本位”观点很符合汉语文化基因进化发展的特点。于是，我与徐教授约定：2000年6月3日，在纪念《马氏文通》发表100周年学术交流会上见面。这之后，从他给我的《语言论——语义型语言的结构原理和研究方法》（东北师范大学出版社）一书中，我了解到徐先生的“字、辞、块、读、句”对明确地区分汉语的文化基因文本元素组合——超子域的几个进阶层式。遗憾的是：那一段时间，徐教授虽然认为我的文章“提出了一些重大的问题”，但是，“因为我们研究的领域不同　”，故无法直接给予支持。后来，我认真读了徐通锵教授给我的《语言学基础理论》（北京大学出版社）、俞士汶教授给我的《计算语言学论文集（4）》（北京大学计算语言研究所）和张全教授给我的《概念层次网络（HNC）》（清华大学出版社）等书的有关论文，对比徐通锵、陆俭明和黄曾阳三位学者的观点，还特别调查了近期国际国内自然语言理解及中文信息处理领域的有关情况，因此，我认为有必要具体地公开我所考虑的如何确立在文化基因工程中对语言发展进阶层式进行划分的标准以及与之相应的协同智能计算语言数据库的构造。

众所周知，由于目前通用计算机中采用的二进制数表示字母、数字、符号以及控制符的美国标准信息交换码，即ASCII，可以说在根本上还不可能直接构造出基于汉语文化的计算机芯片、操作系统和编程语言。同时，由于美国标准信息交换码不表示汉字，所以，建立在ASCII基础之上的汉字信息交换码（GB2312）、中文内码扩展标准（GBK）和基于多八位编码字符集标准（ISO10646）的国家标准（GB13000.1）的中文信息处理的效率，都远不如直接采用英语处理知识信息数据的效率高。

由此可见，现有技术，对计算机处理汉语而言，不仅不是最佳的，而且，还存在根本缺陷或不足。

发明内容

本发明的目的在于提供协同智能计算语言数据库的设计方法，以便于自然语言理解及中文信息处理领域的开发人员设计出效率更高的标准化共享语言知识数据库，也便于用户借助它定制适合自己的个性化独享语言知识数据库，同时，还为设计中的基于文化基因的协同智能计算系统提供便于处理多学科知识信息数据的基础加工平台。

本发明的目的是通过下述技术方案实现的，即：

协同智能计算语言数据库的设计方法，是对“一种知识信息数据处理方法及产品”发明专利说明书和其中公开不充分一项具体技术的改进措施，即：通过建立语言文字的子全域和超子域进阶层式的一系列基础表，构成人机协同对自然语言进行定性分析和定量分析的高效工具平台，它涉及现行的数据库和数据仓库技术以及相应的计算机软、硬件技术产品的直接应用，其特征在于：

首先，把由汉语基本笔画或英语基本字母构成的基础表中的这种元素集合，明确地定义为子全域，分表序号为0，以此作为计量语言文字的基准参照系，同时，因其中的笔画或字母的个数可穷举或实现完全归纳，故在此被明确地定义为基因文本元素，以便计算机复用时进行自动计量；

其次，把语言发展进阶层式各一览表构成的各相应基础表中组合部件的集合，明确地定义为超子域，分表序号为：1、2、3、4、5、6、7、8、9、10、11、12，以此作为计量语言文字的应对参照系，同时，因其中的具体组合部件的个数不可穷举或只能实现相对完全归纳，故在此被明确地定义为基因文本元素组合部件，以便计算机复用时进行自动计量；

最后，在全域数码（a+bi&…）构成的总参照系中，明确地给出各个子全域和超子域各进阶层式一览表的总的统一的通用语言的基础表的id的特定存放序位——由国际及国家的标准化组织认同，在此之前先由用户通过定制各分表的形式由使用单位或有关机构协商选定。

本发明的有益效果在于：既能帮助自然语言理解及中文信息处理领域的开发人员设计出更高效率的标准化共享语言知识数据库，又能帮助普通的广大用户更容易地定制适合自己的个性化独享语言知识数据库，还能为设计中基于文化基因的协同智能计算系统提供一种高效处理多学科知识信息数据的基础加工平台，并且能显著地提高人机协同对语言文字进行定性分析和定量分析的工作效率。

附图说明

图表是协同智能计算语言数据库的设计方案一览表。它以一览表的形式对汉语和英语的子全域和超子域各进阶层式的总说明，其中的内容一目了然，是建立各个具体的基础表的操作指南。

具体实施方式

实施例1是采用微软Office（办公系统软件）的access（存取）数据库的基础表制作的汉语和英语的子全域与超子域进阶层式各基础表的设计说明。由基因文本元素及其组合构件集合成某一语种具体的协同智能计算语言数据库的0、1、2、3、4、5、6、7、8、9、10、11、12个基础表，在各语种中的各个表中的具体成员数目均采用各自相应的的基础表的id形式进行自动计量，汉语、英语和其它语言的各级基础表均如此。

以下结合图1与实施例1对本发明的技术实施方案作进一步说明：

实施例1通过图表的一览表形式把汉语和英语的区别与联系一目了然地呈现在读者面前，以此指导开发者或普通用户进行有针对性的选择，使生成方式与采集方式相结合，从而，高效率地建立语言知识数据库。

开发者或普通用户根据本发明方案仅仅使用access，就很容易有针对性地选择基本笔画、三种偏旁部首、字、辞、语、读、句、段、篇、章、书中的相关部份，建立合乎标准的汉语语言知识数据库或数据仓库。

图1是协同智能计算语言数据库的设计方案一览表。

协同智能计算语言数据库的设计方案一览表

编号	机器序号	分表序号	汉语	拼音	英语	其它
1		0	基本笔画	字母表	26 个字母	
2		1	不成字偏旁部首		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		4	单音节的“字”（独字组）可标：顿号	单音节	单音节的单词	
6		5	复音节的“辞”（复字组）分：离心与向心	复音节	复音节的单词	
7		6	多音节的“语”（多字组）含：两种成份	多音节	多音节的单词	
8		7	标逗号的“读”（表示：语气上的停顿）	标逗号的多音节	词组或短语	
9		8	标句号的“句”（表示：语义上的停顿）			
10		9	须提行的“段”（表示：逻辑上的转换）			
11		10	须题名的“篇”（表示：主题上的区别）			
12		11	须分节编目的“章”（表示：层次的转换）			
13		12	须分类编册的“书”（涉及书库或图书馆）			
	计算语言	自然语言	：（形义结合） 汉	语 （拼音形式）	英语	

（2002年11月应邀在北大、清华、中科院等相关研究部门做专题介绍）

计算语言学专家（清华大学苑春法）**有针对性的评语**之一：

专家评语：

“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。但是同时我也认为这13张表的构建是一件消耗大量人力物力的工作。”

（清华大学计算机科学与技术系**智能技术与系统国家重点实验室**中文信息处理组苑春法教授）

原文附件：

邹晓辉先生：

你好！谢谢你在清华的讲座。

由于时间关系，不能长谈。仅仅从几个小时的讨论交流中对你理论全貌尚未能得到一个清晰的了解。从交谈中，我认识到你的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。但是同时我也认为这13张表的构建是一件消耗大量人力物力的工作。因为仅仅一个汉语的树库的建立就是一件浩繁的工作，至今尚未完成；而它仅仅是你的数据库中的一部分。所以我建议在经过充分酝酿和充分的人力财力准备的基础上再启动这件事。

祝你工作顺利！

苑春法 2003.1.3

计算语言学专家（教育部语言文字应用研究所鲁川）有针对性的评语之二：

专家评语：

“协同智能计算语言数据库的设计方案中的13张表格富有新意。按照这13张表所建立的系统，对于汉语分析中的各种歧义就有可能得到初步的解决。这是一件既有创新意识又极为艰巨的系统工程。这13张表的构建充分体现出你能站在一个较高的起点上善于集中现有各家学派的优点，但是也要看到各家学派所存在的分歧颇似‘冰炭’难以共存，所以还是应以一个学派的理论为主，适当吸收各家之长。因而必须建立一个具有汉语特色的符合知识经济时代需求的新学派、新理论。”

计算语言学专家：鲁川（教育部语言文字应用研究所研究员）

中国中文信息学会计算语言学专业委员会（首届）主任

原文附件：

邹晓辉先生：

你好！感谢你寄来的文件。拜读之后，获益良多。

首先是深为你的精神所感动。你为了我国国民经济和全社会的信息化以及中国计算语言学的发展呕心沥血，可敬可佩！

二十一世纪是知识经济的时代，在知识经济社会中最重要的就是知识。

语言是人类知识的编码系统。所以，许多领域的专家都十分关注语言学的发展。

知识经济时代的特征是“经济全球化”。频繁的国际经济、政治、文化交往导致了蓬勃发展的“第二语言教学”，在我国是外语教学和对外汉语教学。并有海量的翻译任务急需高质量的翻译，这二者都迫切要求对汉语和英语有深刻的符合民族思维模式的突破性研究和对不同民族语言的对比性研究。知识经济时代的支柱产业是信息产业。畅销产品是信息产品(电脑、电视机、多媒体录放机、移动电话、彩色打印机、激光照排机、复读机、快译通……等），这些产品也是“语言产品”，生产这些产品必须依靠先进的科学技术，也包括“语言科学技术”。

从工业时代过渡到信息时代，特别是面对信息时代的高级阶段知识经济时代，人们认识到“语言科学技术也是第一生产力”，仅仅从不算太多的交谈中，我觉得你的协同智能计算语言数据库的设计方案中的13张表格富有新意。按照这13张表所建立的系统，对于汉语分析中的各种歧义就有可能得到

初步的解决。这是一件既有创新意识又极为艰巨的系统工程。

这13张表的构建充分体现出你能站在一个较高的起点上善于集中现有各家学派的优点，但是也要看到各家学派所存在的分歧颇似“冰炭”难以共存，所以还是应以一个学派的理论为主，适当吸收各家之长。因而必须建立一个具有汉语特色的符合知识经济时代需求的新学派、新理论。

从你目前所完成的工作来看，只能说是一个可喜的开端。在2008年北京奥运会和2010年上海世博会之前对“语言产品”有空前的需求，大量的能够创造经济效益的新产品（中国旅游翻译器、对外汉语教学高效软件等）在呼唤着有远见的企业家的投资。建立比陈肇雄公司水平更高的“语言产业”是指日可待的。

得悉你已获得有关方面的大力支持，谨向支持这一重要系统工程到有关领导表示深深的敬意！

我将竭尽全力，对这一有着光明前景和经济效益的语言工程给以最大的帮助。依靠这个语言工程来实现毕生的理想，并跟你们一起报答亲爱的祖国！

有新的进展请尽早告知，我所掌握的若干领先的思路和技术方案都将为这个语言工程服务。附件是将在《语言科学》期刊发表的论文。

祝大家坚持不懈争取最后的胜利！

教 育 部 语 言 文 字 应 用 研 究 所 研究员
中国中文信息学会计算语言学专业委员会首届主任 鲁 川
北 京 大 学 计 算 语 言 学 研 究 所 教 授

专家观点

“当务之急是系统地建设针对大规模真实文本的语言资源库，即经过多级深层次加工的语料库以及语法库、语义库等。这些基础的东西做得不扎实，中文信息处理就很难上一个大台阶。”

（清华大学**孙茂松**教授“谈中文信息处理领域面临的机遇和挑战”）

背景介绍 学术交流

1、牛刀小试

1997年我设计的“多语翻译系统（同义句词及音形转换）”获中国专利技术博览会金奖，当时主要采用的是语法分析和基于规则的方法，辅之以一定规模的熟语料和实例，思路简单清晰、目标具体直接，在受限范围的机译（实质上是有针对性的重用与用户界面优良的机助人译）效果相当好。扩大到非受限范围，由于普通流行文本往往不规范，必须强化语义分析，可是，意义理论至今也都还不成熟，故消歧难题始终存在。这之后，我与陈肇雄（中国电子集团副总裁，南方软件园董事长，华建集团原董事长）、黄河燕（华建集团总裁）和关培忠（译星公司总工程师）有一些交流。

2、再次攀登

2000年5月我设计了“一种知识信息数据处理方法与产品”2001年6月获中国专利技术博览会金奖。这之后，我与张普（北京语言文化大学计算语言学研究所原所长）和林杏光（中国人民大学教授，对外语言文化学院学术委员会副主任，中国中文信息学会理事和学术委员，中国计算语言学专委会专委）进行了一些交流。张普教授认为：该设计是一个大项目，须等待机会展示其实力。林杏光教授认为：融智学理论和设计方法有原创性，建议：1、出版融智学专著，2、开发融智系列产品，3、组织融智团队。这一段时间，我还与北京大学的徐通锵（中文系语言学教研室原主任）、王洪君（中文系语言学教研室主任）、陆俭明（世界汉语教学学会会长，中国语言学会副会长，北京大学人文学部学术/学位委员会委员，北京大学汉语语言学研究中心主任，北京大学计算语言学研究所学术顾问）和俞士汶（北京大学信息科学技术学院教授，计算语言学研究所学术指导委员会主席。兼任中国计算机学会理事和学术委员会副主任、新加坡《汉语语言与计算学报》联合主编等职）以及郭雷（中国科学院院士、中国科学院统科学研究所所长）等有了一些接触和交流。其中，徐教授的字本位汉语理论（见其专著《语言论》1997《基础语言学教程》2000北京大学出版社），给我印象很深。我认为他对汉语与西方语言

的比较是很到位的。四川大学陈雨思副教授从系统科学的角度对该设计作了高度评价。中国人民解放军洛阳外国语学院计算语言学研究室主任易绵竹教授（国际信息化科学院院士，中国中文信息学会理事）来信说“当前语义研究的理论方法还需融合统一，您创立的融智学新范式提炼出协同智能主体的概念体系具有原创性，想必对自然语言语义信息的处理将引发一场革命。”上述交流使我很受鼓舞，我开始更多地关注计算语言学和基础语言学的发展，同时，公开了融智学的部分理论。

3、展示模型

2002年11月我设计了一个可计算、可操作、完全数字化的自然语言理解的总量控制模型（GTCM）--《协同智能计算语言数据库的设计方法（发明专利申请号02153511.6》。这之后，我应邀到北京，参加了几次重要的学术交流（包括在北大、清华、中科院等好几个单位）。其中，在北大计算语言学研究所的交流会使我认识到，在各种汉语语言观之间，没有中间道路可走。随后，在清华大学国家智能实验室（中国科学院院士张拔、孙茂松、苑春法、陈群秀、周强）、中国科学院国家智能实验室（中国科学院院士陆汝黔、曹存根）、微软亚洲研究院自然语言理解组（中国计算语言学专委会主任黄昌宁、周明）、华建集团（中国科学院计算机语言研究中心主任黄河燕、知网创办人董振东）和中软译星（关培忠和他领导的开发团队）交流（重点谈模型的13张表）。陆俭明与胡俊锋明确肯定了我提出的融智这一概念。俞士汶等表示对融智学（支持该模型的理论体系）精髓很感兴趣。鲁川（语言文字应用研究所研究员、北京大学计算语言学研究所兼职教授、中国计算语言学专委会首届主任）对模型的13张表给予了肯定的评价并认为：“这个模型太好了！我遇到了知音。因为，邹晓辉你今天做的事（指由这13张表展示的文化基因工程）就是我鲁川明天和后天想做的事（指汉语基因工程）”。黄昌宁教授与我探讨了这13张表与大脑结构的关系。苑春法教授仔细思考后认为模型的13张表具有原创性，能对汉语实现消歧，同时+

也认为工作量很大。黄河燕教授与我也谈到了消歧和13张表的关系。董振东研究员与我谈了如何实现的事宜。其他各位教授也都表示出了浓厚的兴趣。

第9、10两帖

协同智能计算知识数据库的设计方法

技术领域

本发明属于计算语言和知识工程领域，涉及中文信息处理、自然语言理解、机器翻译、知识获取与重用，进一步是协同智能计算知识数据库的设计方法。

背景技术

“融智学（新范式）”（见系统科学之窗论文专区）、“一种知识信息数据处理方法与产品（00109380.0）”和“协同智能计算语言数据库的设计方法”（02153511.6），虽然公开了子全域与超子域各进阶层式的语言形式化技术方案，但并没有公开已知域和目标域的知识形式化技术方案，特别是协同智能计算知识数据库的设计方法。它们是关于协同智能计算系统如何消歧的科学理论（涉及：信息、智能和理解等基本概念的本质与知识的计量或测度的理论探讨或讨论）与技术实践方案（涉及：中文信息处理、自然语言理解、机器翻译、知识获取与重用等具体领域的技术实践或探索）。

正如清华大学智能技术与系统国家重点实验室苑春法教授所说：“邹晓辉先生：你好！谢谢你在清华的讲座。由于时间关系，不能长谈。仅仅从几个小时的讨论交流中对你理论全貌尚未能得到一个清晰的了解。从交谈中，我认识到你的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。但是同时我也认为这13张表的构建是一件消耗大量人力物力的工作。因为仅仅一个汉语的树库的建立就是一件浩繁的工作，至今尚未完成；而它仅仅是你的数据库中的一部分。

所以我建议在经过充分酝酿和充分的人力财力准备的基础上再启动这件事。祝你工作顺利！苑春法 2003.1.3”

众所周知，计算机如何消歧至今仍是制约中文信息处理、自然语言理解、机器翻译、知识获取与重用等具体技术领域的发展瓶颈。对可扩展标记语言（Extensible Markup Language，XML）也不例外。

目前，其它相关的现有技术方法及产品，至多涉及图1语言文字表所述序号为0、1、2、3、4、5、6、7、8、9、10、11、12 进阶的某些部份。

例如：普林斯顿大学的词网（WordNet）、中科院计算机语言信息工程研究中心研究室（董振东和董强等）的中英文双语词网（HowNet）、北京大学计算语言学研究所（于江生等）的中文概念辞书（Chinese Concept Dictionary，CCD）和清华大学智能技术与系统国家重点实验室（周强等）的汉语树库都是语汇层面的网络电子版概念词典。

电子图书浏览和联机帮助虽是篇章层面的网络电子版图书，例如：IBM（国际商业机器）公司的数字化图书馆和Microsoft（微软）公司的数字化百科全书以及企业级的知识管理，尽管它们在后台数据库或数据仓库与前台交互界面以及计算机网络构成的软硬件支持环境的共享或重用（特别是检索查询）方面的效率很高，但是，就语言文字和学科知识的处理而言，在实质性理论和相应的工程化技术方面，仍然面临计算机如何消歧的难题，可以说它们只初步解决了知识文本的粗加工问题，而对学科知识本身怎样有效地进行深入处理（在内容方面）或精加工（在形式方面）的问题还处于探讨阶段，因为，不仅计算机如何消歧的难题至今仍未解决，而且，自然人如何消歧的难题至今也仍未彻底解决，否则，科学家们就不会继续在科学理论方面进行对信息、智能和理解等基本概念的本质与知识的计量或测度的理论探讨或讨论，也没有必要继续在技术实践方面进行对中文信息处理、自然语言理解、机器翻译、知识获取与重用等具体领域的技术实践或探索。

发明内容

本发明的目的是提供协同智能计算知识数据库的设计方法，既便于开发人员设计效率更高的标准化共享知识数据库，又便于用户借助它定制适合自己的个性化独享知识数据库，更便于协同智能计算系统用它做多学科知识信息数据处理的专业基础加工平台。

本发明的目的是通过下述技术方案具体实现的，即：

协同智能计算知识数据库的设计方法，是协同智能计算语言数据库的设计方法（02153511.6）进一步发展或应用的产物，即：在语言文字表（图1）的基础之上建立学科知识表（图2）、规范知识表（图3）和直观知识表（图4），其特征是：从子全域与超子域各进阶层式中选出已知域和目标域，构成人机协同对学科知识进行定性分析和定量分析的专业基础加工平台，其中，

已知域，是关于学科分支及课题的知识点的集合；

目标域，是关于“问”与“答”的知识点的集合。

知识点是以下分类的规范化知识表达（语言文字或符号的基本组合）：

1、事实，着重客观的记录与真实的再现；

2、规律，强调语言的准确与数学的精练含公式的简明和图表的直观以及限制条件的清楚明白；

3、原理，强调对客观机理的系统说明和对其必然性与重要性及现实性或可行性的完整论述；

4、例题，着重方法的理性步骤与示例的感性操作；

5、习题，强调必要的重复与形式的变换；

6、试题，突出灵活的应用与积极的应对；

7、简纲，要求简明扼要；

8、详纲，要求系统周全；

9、反例，突出有理有据的经典实例。

相对完全归纳的已知域是识别、理解和表达的基础，具有明确针对性的目标域是识别、理解和表达的重点或焦点。

本发明的作用是：

提高人机协同对学科知识进行定性分析和定量分析的工作效率，

为推广符合终身教育观念的产、学、研、用、算一体化的生产式教学法提供协同智能化的专业基础加工平台。

其有益效果还在于：

它不仅能帮助开发人员设计出更高效率的标准化共享协同智能计算知识数据库，也便于用户借助它定制适合自己的个性化独享协同智能计算知识数据库。

一旦基于服务器的共享协同智能计算知识数据库和基于终端的独享协同智能计算知识数据库与相应的软硬件有机地组合在一起，就可十分方便地构建出基于文化基因或全域数码的高效的协同智能计算系统（包括协同智能计算机和协同智能计算网）。

附图说明

图1是协同智能计算语言数据库的设计方案一览表（语言文字表）。

图2是协同智能计算知识数据库的设计方案一览表（学科知识表）。

图3是协同智能计算知识数据库的设计方案一览表（规范知识表）。

图4是协同智能计算知识数据库的设计方案一览表（直观知识表）。

具体实施方式

实施例1与图1、图2、图3和图4的一览表是采用微软Office（办公系统软件）的access（存取）数据库的基础表制作的汉语和英语的子全域与超子域进阶层式以及从中筛选收敛集合而成的已知域和目标域的一系列基础表的设计说明。本发明的技术方案就是：在语言文字表（图1）的基础之上建立的学科知识表（图2）、规范知识表（图3）和直观知识表（图4）。

协同智能计算知识数据库中学科知识表（图2）、规范知识表（图3）和直观知识表（图4）等一览表的各科各级的基础表，都是由基因文本元素及其组合构件集合而成的某语种的协同智能计算语言数据库的0、1、2、3、4、5、6、7、8、9、10、11、12进阶的基础表的子集，各分科知识表的具体成员数目也都采用各基础表的自动编号（id）形式进行自动计量，不仅汉语、英语和其它语言的各级基础表均如此，而且，各科知识的基础表也如此。

以下结合附图与实施例1对本发明技术方案作进一步说明：

本发明及其实施例通过图1、图2、图3和图4的一览表形式把汉语和英语的区别与联系一目了然地呈现在读者面前，以此指导开发者或普通用户进行有针对性的筛选收敛集合——知识获取。

首先，根据基因文本的进阶层式，语言文字知识获取的基本方式有：

a、生成法，由低到高，逐级合成；

b、采集法，由高到低，逐级分解；

c、混合法，针对需要，跨级插入。

其中获取的语言文字知识，既含离散的普通常识，也含系统的学科知识。

接着，对具体的常识与学科知识进行科学划分和属性标注，为去冗存要奠定自动化处理的基础。

然后，应用access（存取）数据库的选定内容筛选功能，相对完全地有针对性地筛选收敛集合0、1、2、3、4、5、6、7、8、9、10、11、12进阶的相关部份，建立相对完全归纳的已知域标准化共享知识数据库和具有明确针对性的目标域个性化独享知识数据库，其中，各终端的access表与服务器中立方体（cub in SQL server）的access表的自动编号（id）在相应的数据库或数据仓库中的全域数码（a + bi &…）都有其具体的惟一的序位。

实施例1通过图2、图3和图4的一览表把离散的知识点和系统的知识框架以语言文字和多媒体形式一目了然地呈现在读者面前，以便于指导开发者或普通用户通过相对完全地有针对性地筛选收敛集合建立分科知识表。

具体地说，本发明及其实施例1的具体实施方法可以详述如下：

一、基本步骤：

1、设置知识表的记录与属性——定义数据结构和分配文件空间，以文化（符号）和物化（物理）

的形式构造各种各样的表和立方体，

具体由语言表标注、知识表规范和物象表链接三个环节组成，

2、导入和使用知识信息数据——填充知识点的形式化数据——规范化知识表达的语言文字或符号的基本组合，并以此为基础制作相应的查询、窗体、报表、页、宏、模块，

其特征在于：

根据从子全域与超子域各进阶层式中选出的已知域或目标域，设置知识表的属性——列，即：在语言文字表的分表序号为0、1、2、3、4、5、6、7、8、9、10、11、12 进阶的十三类子表中，设置学科属性——列，选定内容筛选记录——行，筛选收敛集合构成一系列分科知识表。

二、产品形式：

1、基础表类型，采用分离形式的学科知识表、规范知识表和直观知识表三种形式，其整合形式构成协同智能计算知识数据库，

2、产品的功能形式，采用教具、学具、玩具、用具、工具五种形式，

3、产品的载体形式，采用电子数字出版物（如：芯片、光磁盘、终端、服务器、网络等形式）与传统印刷出版物及其协同互补的各种形式，

其中，电子数字出版物是标准化与个性化统一的高效工具，而传统印刷出版物的产品形式的形成，则有赖于从电子数字出版物的学科知识表、规范知识表和直观知识表中选择最适合采用传统印刷出版物的活页卡、活动表、书本、书刊和手册的部份。

三、使用方式：

设计者或普通用户与计算机及其网络之间的人机协同，涉及：

1、语言文字、学科知识和直观物象等处理对象，

2、事实、规律、原理、例题、习题、试题、简纲、详纲、反例等形式，

3、听、说、唱、读、写、译等交互方式，

4、数、字、图、表、音、像、立体、活体八大形式体系，

5、产、学、研、用、算五大功能系列，

6、定性分析、定量分析、结构分析、程序分析、定向分析、定位分析等六大分析方法，

7、自然物和人工物在虚（着重感觉方面）与实（着重行为方面）两方面的知行关系，

8、语言文字数据库和分科知识数据库的使用与生产，

9、围绕各表的属性和记录的自动编号（id）的整合，构成在相应的数据库或数据仓库中的全域数码（a + bi &…）的序位关系，涉及知识表达或形式化知识的供、产、消关系。

四、语言表标注，学科知识表（图2）的构成方法：

1、在语言文字表的十三类分表的具体语种表中，设置一系列用于标识学科的属性——列，选定内容筛选记录——行，筛选收敛集合构成一系列学科知识子表，

2、在具体的学科知识子表中，设置一系列用于标识学科分支的属性——列，选定内容筛选记录——行，进一步筛选收敛集合构成一系列学科分支知识孙表，

3、构成已知域学科知识表——学科知识子表和学科分支知识孙表，

4、设置一系列用于标识课题的属性——列，选定内容筛选记录——行，再进一步筛选收敛集合构成目标域学科知识表。

五、知识表规范，规范知识表（图3）的构成方法：

1、在学科知识表中，设置对应序号的属性——列，

2、同时，设置事实、规律、原理、例题、习题、试题、简纲、详纲、反例等规范化知识表达的属性——列，

3、选定内容筛选记录——行，筛选收敛集合构成规范知识表，包括各学科规范化知识表达的事实表、规律表、原理表、例题表、习题表、试题表、简纲表、详纲表、反例表。

六、物象表链接，直观知识表（图4）的构成方法：

在学科知识表或规范知识表中，设置数、字、图、表、音、像、立体、活体的属性——列，选定内容筛选记录——行，筛选收敛集合构成相应的数、字、图、表、音、像、立体、活体等一系列超级链接的直观知识表。

总之，图2、图3和图4等一览表是建立分科知识表的操作指南。

协同智能计算语言数据库的设计方案（**语言文字表**）												
编号	进阶	汉语	拼音	英语	法语	德语	俄语	日语	西班牙语	葡萄牙语	等等	其它语种
1	0	基本笔画	字母表	字母表	表 1	表 1	表 1	音图	表 1	表 1	表 1	表 1
2	1	不成字组合		词头和词尾	表 2	表 2	表 2	表 2	表 2	表 2	表 2	表 2
3	2	变形字组合		前缀和后缀	表 3	表 3	表 3	表 3	表 3	表 3	表 3	表 3
4	3	字中字组合		词根	表 4	表 4	表 4	表 4	表 4	表 4	表 4	表 4
5	4	单音节的字	单音节	单音节词	表 5	表 5	表 5	表 5	表 5	表 5	表 5	表 5
6	5	复音节的辞	复音节	复音节词	表 6	表 6	表 6	表 6	表 6	表 6	表 6	表 6
7	6	多音节的语	多音节	多音节词	表 7	表 7	表 7	表 7	表 7	表 7	表 7	表 7
8	7	标逗号的读	标逗号	词组或短语	表 8	表 8	表 8	表 8	表 8	表 8	表 8	表 8
9	8	标句号的句		标句号的句	表 9	表 9	表 9	表 9	表 9	表 9	表 9	表 9
10	9	须提行的段		须提行的段	表 10	表 10	表 10	表 10	表 10	表 10	表 10	表 10
11	10	须题名的篇		须题名的篇	表 11	表 11	表 11	表 11	表 11	表 11	表 11	表 11
12	11	须分节的章		须分节的章	表 12	表 12	表 12	表 12	表 12	表 12	表 12	表 12
13	12	须分册的书		须分册的书	表 13	表 13	表 13	表 13	表 13	表 13	表 13	表 13

图 1

协同智能计算知识数据库的设计方案（1 **学科知识表**）													
编号	进阶	汉语	拼音	英语	序号	符号	语言	数学	物理	化学	生物	等等	其它学科
1	0	基本笔画	字母表	字母表	0-0	表 1	表 1	表 1					
2	1	不成字组合		词头和词尾	1-1	表 2	表 2	表 2					
3	2	变形字组合		前缀和后缀	2-2	表 3	表 3	表 3					
4	3	字中字组合		词根	3-3	表 4	表 4	表 4					
5	4	单音节的字	单音节	单音节词	4-4	表 5	表 5	表 5	表 5	表 5	表 5	表 5	表 5
6	5	复音节的辞	复音节	复音节词	5-5	表 6	表 6	表 6	表 6	表 6	表 6	表 6	表 6
7	6	多音节的语	多音节	多音节词	6-6	表 7	表 7	表 7	表 7	表 7	表 7	表 7	表 7
8	7	标逗号的读	标逗号	词组或短语	7-7	表 8	表 8	表 8	表 8	表 8	表 8	表 8	表 8
9	8	标句号的句		标句号的句	8-8	表 9	表 9	表 9	表 9	表 9	表 9	表 9	表 9
10	9	须提行的段		须提行的段	9-9	表 10	表 10	表 10	表 10	表 10	表 10	表 10	表 10
11	10	须题名的篇		须题名的篇	10-10	表 11	表 11	表 11	表 11	表 11	表 11	表 11	表 11
12	11	须分节的章		须分节的章	11-11	表 12	表 12	表 12	表 12	表 12	表 12	表 12	表 12
13	12	须分册的书		须分册的书	12-12	表 13	表 13	表 13	表 13	表 13	表 13	表 13	表 13

图 2

协同智能计算知识数据库的设计方案（2 **规范知识表**）														
编号	进阶	汉语	拼音	英语	序号	事实	规律	原理	例题	习题	试题	简纲	详纲	反例
1	0	基本笔画	字母表	字母表	0-0									
2	1	不成字组合		词头和词尾	1-1									
3	2	变形字组合		前缀和后缀	2-2									
4	3	字中字组合		词根	3-3									
5	4	单音节的字	单音节	单音节词	4-4							4		4
6	5	复音节的辞	复音节	复音节词	5-5							5		5
7	6	多音节的语	多音节	多音节词	6-6							6		6
8	7	标逗号的读	标逗号	词组或短语	7-7								7	7
9	8	标句号的句		标句号的句	8-8	8	8	8	8	8	8		8	8
10	9	须提行的段		须提行的段	9-9	9	9	9	9	9	9		9	9
11	10	须题名的篇		须题名的篇	10-10	10	10	10	10	10	10			10
12	11	须分节的章		须分节的章	11-11									
13	12	须分册的书		须分册的书	12-12									

图 3

协同智能计算知识数据库的设计方案（3 **直观知识表**）													
编号	进阶	汉语	拼音	英语	序号	数	字	图	表	音	像	立体	活体
1	0	基本笔画	字母表	字母表	0-0	表 1	表 1	表 1	表 1	表 1	表 1	表 1	表 1
2	1	不成字组合		词头和词尾	1-1	表 2	表 2	表 2	表 2	表 2	表 2	表 2	表 2
3	2	变形字组合		前缀和后缀	2-2	表 3	表 3	表 3	表 3	表 3	表 3	表 3	表 3
4	3	字中字组合		词根	3-3	表 4	表 4	表 4	表 4	表 4	表 4	表 4	表 4
5	4	单音节的字	单音节	单音节词	4-4	表 5	表 5	表 5	表 5	表 5	表 5	表 5	表 5
6	5	复音节的辞	复音节	复音节词	5-5	表 6	表 6	表 6	表 6	表 6	表 6	表 6	表 6
7	6	多音节的语	多音节	多音节词	6-6	表 7	表 7	表 7	表 7	表 7	表 7	表 7	表 7
8	7	标逗号的读	标逗号	词组或短语	7-7	表 8	表 8	表 8	表 8	表 8	表 8	表 8	表 8
9	8	标句号的句		标句号的句	8-8	表 9	表 9	表 9	表 9	表 9	表 9	表 9	表 9
10	9	须提行的段		须提行的段	9-9	表 10	表 10	表 10	表 10	表 10	表 10	表 10	表 10
11	10	须题名的篇		须题名的篇	10-10	表 11	表 11	表 11	表 11	表 11	表 11	表 11	表 11
12	11	须分节的章		须分节的章	11-11	表 12	表 12	表 12	表 12	表 12	表 12	表 12	表 12
13	12	须分册的书		须分册的书	12-12	表 13	表 13	表 13	表 13	表 13	表 13	表 13	表 13

图 4

10004 2002.

5th Chinese Lexical Semantics Workshop (CLSW-5新加坡)

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

摘要：语义词汇典例（SVDE）的总量控制模型，既是一种新理论，又是一种新方法，还是一种新工具。在人机协作网络(融智系统)中有两种总量控制模型,即:关于自然语言理解的文本总量控制模型(GTCM)和音节总量控制模型（GSCM）。 GTCM表示在GLPS中的文本分为0-16个进阶。GSCM表示在SVDE中的音节分为1-n个进阶。SVDE的义项由成对的编号序列控制。字与解释字的义项的字组之间遵循1对n的法则构成母语的SVDE（单语义项字典）。无论基于并列性还是基于合成性双语的观点，解释字的义项的汉语字组与解释词的义项的英语词语之间遵循1对1的法则构成双语的SVDE（双语用例词典）。

关键词：字组细分、总量控制、人机协作、单语义项字典、双语用例词典

一、绪言

在人机协作网络（融智[1]系统）中有两种总量控制模型，即：关于自然语言理解的文本总量控制模型（GTCM）和音节总量控制模型（GSCM）。本文探讨词汇一级的模型GSCM，属于计算语言学分支汉语词汇语义学的课题，位于全球语言定位系统（GLPS）与全球知识定位系统（GKPS）的结合部 [2]。SVDE处理一字多义的方式，与学科内流行的“贴标签[3]”的方式不同，算得上是一种高效处理词汇语义的简便方法。为进一步寻找消除自然语言理解的语义障碍[4]的新途径，本研究的侧重点，不是“埋头拉车”，如：“贴标签”或分析“素、类、槽、格[5][6]”，而是“抬头看路”，如：把握形式化的方向、辨别可否计算、考虑知识表达以及关注各种本位说[7][8]。为了让读者以小见大、窥斑知豹，本文从理论回顾、模型发展、个案分析三个方面进行综述，然后，介绍方法、结果和结论。VSDE涉及两个假设：1、“单独存储(并列性双语者)与共同存储（合成性双语者）”[9]可由融智系统整合为典与例。2、汉语的混音节线串型字组（词语）是单音节层面型字组（汉字）与英语的混音节词语之间无歧义连接（同意并列）的纽带 （旨在保证双语的义项对译）。本课题的贡献在于：1、提出了字组细分的观点和GSCM；2、提出了对语言和知识进行直接表达和间接计算的策略（区别于间接表达和直接计算的策略）；3、为推行“产、学、研、用、算”一体化的人机协作实施语文系统工程和知识系统工程提供了SVDE体系，对消除词汇一级的形式歧义和内容歧义十分有益。

二、综述

对语义问题的认识有一个由简单到复杂再由复杂到简单的过程。下面，分三个方面进行综述。

理论回顾

1、在**语言哲学**方面，与本研究密切相关的是“意义”问题。“意义、词语、事物”这个“语义三角”一直是有争议的。以往理论主张“意义”作为词语或概念是不可分的。本研究认为可分。

2、在**语言理论**方面，与本研究密切相关的是“本位”问题，各种本位说，首先反映不同的汉语**语言观**， 其次必然带来相应的各种理论，最后也必然影响汉语理论实践的各个方面。本研究的语言观：1、就古代汉语和现代汉语中与传统一脉相承的语言现象而言，认同字本位。2、就现代汉语中吸收西方语言而发生显著改变的语言现象而言，主张**字组细分**。这样，不仅能较好地与英语等西方语言的词、词组和短语**对译**，而且也能与其他本位观之间建立相互**兼容**的实用接口。本研究认为：汉语是**拼字音节**，汉语的混音节**线串型字组**是单音节汉字与英语的混音节词语之间**无歧义连接的纽带**。本研究为研

究**双语或多语**和一文双语（涉及拼音）[10]以及一文多语（涉及方言）提供了**参照模型**。

3、在逻辑理论方面，与本研究密切相关的是“消歧”的问题。本研究认为：逻辑学实质上是一门研究消歧的学问。例如：二值逻辑和三值逻辑，就是处理二歧性与三歧性的问题。

4、在数学理论方面，与本研究密切相关的是“多元数”问题。本研究认为，“多元数”不仅是数学与逻辑学之间的一个结合点（属下一步的研究课题），而且对复杂性系统的表达特别有用。

5、在认知理论方面，与本研究密切相关的是“双语存储”问题。本研究认为：双语的单独存储模型与共同存储模型各持一端，故提出融智系统的整合协同存储模型（属下一步的研究课题）。

6、在计算语言方面，与本研究密切相关的是“语言的计算与表达”问题。本研究认为：基于规则、统计、实例的处理既可以有**直接计算**（属性值）和**间接表达**（属性标注），也可以有**间接计算**（数字）和**直接表达**（直接呈现母语表达的知识）。词网（wordnet）与SVDE可以兼容。

7、在知识工程方面，与本研究密切相关的是“知识的计算与表达”问题。基于数据库及数据仓库，SVDE可以有效地处理常识性知识，在一定数量或规模的范围以内很有效。

8、在信息理论方面，与本研究密切相关的是“信息的本质”问题。本研究对这个问题的探讨，是与前面的“义”的义项的研究联系在一起的，涉及一般科学的信息定义。

9、在软件理论方面，与本研究密切相关的是“软件的计算与表达”问题。可用程序语言的冗余度很大。这增加了人们对软件编程的神秘感。本研究的方法有利于软件开发以简驾繁、去冗存要。

10、在人工智能方面，与本研究密切相关的是“智能的本质”问题。从时间顺序看，人类智能、人工智能、协同智能，前两者是后者的基础，“理解”属于“智能”体现的一种具体类型。

由此可见，要思考解决语义问题的方案，必然涉及很宽的领域。某个学科认为非常难的问题，在多个学科的角度看来，也许只是“小菜一碟”！

模型发展

本文所述自然语言理解的总量控制模型（GCM）分为文本总量控制模型（GTCM）和音节总量控制模型（GSCM）。其中，在词汇一级，GTCM有1~7七个进阶（即0~6七个表），GSCM有拼音文字（如英语）与非拼音文字（如汉语）的区别，以音节为单位，考虑语义，英语涉及六个进阶，词素是单音节，词、词组和短语都是混音节；汉语涉及三个进阶，字是单音节，字辞语统称字组（其中字视为独字组）是混音节。从字组细分的观点看拼字就是拼音节。上述分析与下表的思想一致。

进阶	机码	表号	汉语	拼音	英语
1		0	基本笔画	字母表	26 个字母
2		1	不成字偏旁部首		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		4	单音节字（独字组）是汉语的基本语言单位基本语	单音节	单音节的词
6		5	双音节字组（双字组）可区分：离心与向心	双音节	双音节的词或语
7		6	多音节字组（多字组）含 4 与 5 两种基本成份	多音节	多音节的词或语

个案分析

汉语词义消歧的文献[11]谈如何标注“看”的词条。转述与分析：1、把“看”视为词，2、对多义的处理（有几个义项就列几行），3、分列标上“词类、义项、主体、客体、英语单词”等属性，4、汉译英时，根据搭配特征，选择与相应词条对应的英语单词，即：“see， watch， read”。显然，“看”与“see， watch， read”之间是“1对3”的关系。从“1”到“3”的转换靠“属性标签”间接实现的。识别或计算的也是“属性标签”或其搭配“特征集”。于是，存在几个问题：a、对汉语义

项形式的表达是间接的。b、对汉语义项知识的计算是直接的。c、汉字与英词的对译出现脱节（绕了一个大弯）。当然，补上相应的汉语字组（义项用例）对译脱节的问题也就迎刃而解了。单从“拉车”的角度看，这个问题似乎很容易解决。遗憾的是，由于“埋头拉车”而没有“抬头看路”，所以，发现不了这个问题的存在（视而不见）。“看路”的人关注的方向不同。思路受制于观点——不同的语言观（大前提 ）导向不同的方向。加上“拉车和看路”的人被习惯所左右——不同策略（小前提）制约选取知识表达和计算的方式。这就失去了改变的可能。本来看似简单的问题也就变得复杂了。就本例而言，就是在“看”（汉字）与“see， watch， read”（英词）之间增加“看见、观看、阅读”（汉语字组）。这样，“1对3”就直接转化为3个“1对1”（SVDE把这个转化一般化，GSCM使其总量可控——注：由于CLSW5对论文页数的限制，显而易见的对照表在此省略），汉译英的歧义自然消除。补上“看”字义项的字组用例并不难。但要改变语言观就非常困难，要改变习惯也不容易。

三、方法

汉语语义词汇典例（VSDE）的总量控制模型（GCM），是根据字组细分的观点和拼字音节的分划方法，把所有的汉语字组（词语）以单音节的字作为汉语的基本语言单位进行计量和排序（1-n）。其中，n表示自然数。也就是说，GSCM表示在VSDE中的音节分为1-n个进阶（由n个表记录）。SVDE的义项由成对的编号序列控制。字与解释字的义项的字组之间遵循1对n的法则构成单语义项字典。无论是基于并列性还是基于合成性双语的观点，解释字的义项的汉语字组与解释词的义项的英语词语之间都遵循1对1的法则构成双语用例词典。GTCM表示在GLPS中的文本分为0~16个进阶。

1、制作SVDE的定性方法——义项字典与双语用例的相互关系a、单语义项字典（参考认知理论的独立存储学说，以汉释汉为例）汉语：从单音节的汉字到混音节的汉语字组，一字多义的义项表述形式为“1对n”；b、双语用例对译（参考认知理论的共同存储学说，以汉译英为例）双语：从汉语的混音节字组到英语的混音节词语，对译的双语用例的表述形式为“1对1”。

2、制作SVDE的定量模型——汉语词汇与英语词汇的进阶层式a、文献《协同智能计算语言数据库的设计方法》曾经把汉语的“字、辞、语”分别排在GTCM的“第4、第5、第6”三个发展进阶层式的位置。b、对词汇一级而言，上述安排过于粗放，故进一步提出从单音节的汉字到混音节的汉语字组的细分方案，并且按照“单字、双字、三字、四字 . . . 多字”（具有可计算性）的表述形式，抽象地采用自然数进行表示，由于考虑到“单音节的汉字”位于GTCM的“第4”这个特定的发展进阶层式的位置，把“字、辞、语”即“第4、第5、第6”三个进阶以内的所有词语合并到一起，再另行按照“单字、双字、三字、四字 . . . 多字”（具有可计算性）的顺序，细分为“0-n”个进阶，并以此命名为：GSCM。c、即：进阶层式数据库的“第4”表或字组细分数据库的第“1”表——单音节的汉字总表，字组细分数据库的第“0-n”表——混音节的汉语字组总表。d、词汇义项典例，即：“1对n”单语解释字典和“1对1”双语对译词语（用例），是前述定性部分的内容。

综上所述，词汇义项典例的总量控制模型（GSCM），由定性和定量两部分构成。

四、结论

词汇义项典例的总量控制模型（GSCM），既是一种新理论，又是一种新方法，还是一种新工具。

1、字组细分的基本观点（GSCM体现的科学原理之一）

在词汇一级，主张对汉语词汇从单音节的字到混音节的字组进行细分的观点。字组细分可使汉语中蕴藏的通用原理更容易显现出来，例如：a、从音节与汉字一一对应的关系来看，汉语是最规范的（如：英语的词的音节就不规范，表现为混音节）。b、字组中字数的增加与语义中项数的减少之间表现出反变关系。c、可按音节数估算汉语使用过程中概念的个数与被重用次数。d、汉语词汇的基本数量，从一字到二字区别大呈上升趋势，再从二字到多字区别大呈下降趋势，例如：字（只有几万个）、二字组（已有人从语料中采集统计出几十万个）、三字组（例如三字经等常用三字词语的数量也不过几万个）、四字组（例如成语只有二万多个）. . . 多字组（如歇后语等常用多字词语的数量则更少）。

2、词汇语义的处理法则（GSCM体现的科学原理之二）

基于字组细分的观点可归纳出自然语言（词汇一级）处理的基本法则。a、词汇义项的定性分析与重用法则（1）单语解释，遵循：“1对n”法则。例如：单音节汉字总表中的汉字编号与义项编号之间，就遵循“1对n”法则。（2）双语对译，遵循：“1对1”法则。例如：混音节字组总表中的字组编号与双语用例编号之间，就遵循“1对1”法则。采用混音节汉语字组表达的义项解释用例的编号与采用其它自然语言语种的词语表示的同一义项解释用例的编号是一致的和通用的。也就是说，尽管计算机前台展示的界面是多样化的，但后台数据库中存储的同一义项解释双语（多语）用例的编号是一致的。b、词汇义项的定量分析与重用法则（1）单音节的汉字总表，包含的汉字编号与义项编号两组数据是不对称的。混音节的汉语字组总表的义项编号与用例编号两组数据（“+”）是一致的，对译的混音节词语总表的义项编号与用例编号两组数据（“-”）也是一致的，而且“+”与“-”是对称的。（2）对符号的计量与重用以字的编号为基准；对语义的计量与重用以义项编号为基准。

五、评论

1、实践意义 为人们汇编义项字典和用例大全，提供了简明的基本操作规范。不仅方便专家而且也方便大众（可共同参与），从而能够汇编一部有史以来规模最大、质量最高、通用性最强或适用面最广的网络版（汉释汉）义项字典与（汉译英）用例大全（其它语种可以此为样板）。在这个基础之上可以很方便地定制各种具体的有明确针对性的出版物（包括：印刷版、电子版、数字版）。

2、理论意义 基于本研究提出的汉语字组细分的观点，不仅发展了汉语字本位的传统，同时也兼容了受外语影响而产生的词语观，而且还可使各种本位说从中找准自己的位置——既不夸大也不缩小，这样，既有利于汉语体系的建立，又便于与世界其它语言体系之间达成较好的交流、沟通与融合。

六、总结

综上所述，GSCM是可计算、可操作、完全数字化的。VSDE，规则简明，提供了大众（广大师生）参与的条件（简单可行有法可依——便于“学法、立法、守法、执法、司法、监督”）。其他任何一种（太复杂、小作坊、各自为政）方法都做不到。本方法充分考虑到了“产、学、研、用、算”一体化的人机协作，不仅为大规模开发各种典与例提供了捷径，也为一个民族或一个国家极大地开发现有的智力资源提供了基础。以往的作法，要么过于依靠机器（“算”），要么过于依靠专家（“研”）。

参考文献

1、邹晓辉《融智学纲要》2004年 见
2、邹晓辉《协同智能计算语言数据库的设计方法》2002年 见http://culturegene.icpcn.com
3、詹卫东《80年代以来汉语信息处理研究述评》见俞士汶等编《计算语言学文集》（第四集）
4、俞士汶《汉字和汉语民族语言进入信息系统》见俞士汶等编《计算语言学文集》（第四集）
5、林杏光《词汇语言学和计算语言学》1999语文出版社年
6、鲁川《汉语语法的意合网络》2001商务印书馆
7、徐通锵《语言论》1997东北师范大学出版社《基础语言学教程》2000北京大学出版社
8、陆俭明、郭锐《汉语语法研究面临的挑战》见俞士汶等编《计算语言学文集》（第四集）
9、汪安圣等《认知心理学》1996北京大学出版社
10、王开杨《“一语双文”的理论基础和面临的困难》见苏培成等编《语文现代化论文集》2002商务印书馆
11、王惠《汉英机器翻译中基于大型语义词典的汉语词义消歧》见黄河燕主编《机器翻译研究进展》2002电子工业出版社
12、俞士汶等编《计算语言学文集》（第四集）见

背景知识：

义项语汇典例（SVDE）的总量控制模型，把中文信息处理的“字处理平台、词处理平台和句处理平台这3个层次”有机地联系在一起，从而，为“中文自动分词和词性自动标注系统”与“其他深层次的语言处理技术，如名词短语捆绑、句法分析、语义分析等”奠定了坚实的基础。它对“面向Internet的文本信息检索、过滤、分类、摘要”，“Internet环境下的机器翻译”，“语音识别”和“大规模的文本挖掘”等领域，都具有非常实用的价值。

孙茂松教授认为“目前最具现实性和可能性的语言处理技术或者说本身研究相对成熟、潜在应用最广泛的技术，非中文自动分词和词性自动标注系统莫属。”

（“谈中文信息处理领域面临的机遇和挑战”）

论汉语字组的细分（重点分析字和字组的关系）

--字组细分方法及其计算模型对汉语字本位理论的意义

关键词语：字本位、字组细分、总量控制、人机协作

论文纲要：

一、绪言

1、领域： 本课题涉及语汇分析方法（基础语言学）和语义词典模型（计算语言学）的交叉学科分支领域。

2、特殊性：通过人机协作，对汉语字组，进行总量可控的细分，融合了崭新的语言观和知识观。

3、重要性：汉语字组细分方法与总量控制模型，不仅在实践方面为后续的语文、知识和软件三大系统工程的顺利开展奠定了语汇一级汉语资源库建设的基础，而且还在理论方面为汉语字本位的观点和体系的推广及理论本身的进一步完善提供了支持。

4、研究途径：通过人机协作，对汉语字组，进行总量可控的细分，分为两个基本途径（或方面）：一方面是以人为主导的研究开发型的义项细分途径（着重本义的信息分析）；另一方面是以机为主导的共享重用型的字符细分途径（着重文本的信息分析）。

5、局限性：本课题只涉及汉语字组的细分--重点分析字和字组的关系。

6、关键假设：研究开发型的超级用户群体拥有的语言技能与科学知识远远多于共享重用型的普通用户群体拥有的语言技能与科学知识。

7、贡献：通过建立汉语的字组细分理论及总量控制模型，开辟了人机协作的新方向。

二、综述

1、字本位的汉语语言观（基础语言学）

以下着重介绍字本位理论提出前后汉语语言观的发展。

a、字本位理论提出前

《马氏文通》引入外语（拼音文字）的词本位语言观。

《转换语法》引入英语（拼音文字）的语本位语言观。

注：语本位，即：短语本位（它等价于：句本位）。

b、字本位理论提出后

《语言论--语义型语言的结构原理和研究方法》

一方面，指出英语（拼音文字）的复本位语言观；

另方面，确立汉语（方块文字）的单本位语言观。

注：复本位，指：词与句。单本位，指：字本位。

2、字组细分的汉语信息处理的知识观或方法论（计算语言学）

a、计算语言学通常直接采用基础语言学的语言观。

b、《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 》提出字组细分的计算语言观。

注：本文将对字组细分的计算语言观进行深入分析。

三、方法

1、数学方法

a、 代数方程

（1）根据汉语字组细分的计算语言观提出自然语言理解的总量控制模型；

（2）根据英语词组细分的计算语言观提出自然语言理解的总量控制模型。

注： 汉语与英语的上述模型，虽然相同，但各自的符号形式体系却不同。

b、 函数表格

注： 由于汉语与英语的数学模型相同，所以理解它们的函数表格也相同。

2、数据处理方法

a、数据库方法

b、数据仓库方法

注： 对汉语与英语采用直接表达与间接计算的策略，因此，对自然语言理解的方式则是通过对数字的直接计算和间接表达的形式实现的，也就是说，采用字组细分的计算语言观可以直接借用现成的数据库以及数据仓库的技术和管理的方法，支持人机协同进行自然语言理解和科学知识表达以及形式信息识别甚至融智软件处理。

3、系统工程方法

a、语文系统工程方法

基于自然语言理解的总量控制模型。

b、知识系统工程方法

基于科学知识表达的结构控制模型。

c、软件系统工程方法

基于形式信息识别的质量控制模型；

基于融智软件处理的质量控制模型。

4、字组分合方法

a、 字组细分方法（义项细分途径）

基于义项细分的科学知识表达，着重“本义”的信息处理，以自然人为主导，突出研究开发的原创性。

b、 组字重用方法（字符细分途径）

基于字组细分的自然语言理解，着重“文本”的信息处理，以计算机为主导，突出学习重用的共享性。

四、结果

1、数学表达方式

c、 代数方程（具体的方程和实例介绍）

d、 函数表格（具体的表格和实例介绍）

2、数据处理形式

a、汉语义项字典（数据库与数据仓库的部分）

b、汉语用例大全（数据库与数据仓库的部分）

3、 系统工程蓝图

a、语文系统工程蓝图（分为：基础教育与高等教育两个阶段）

b、知识系统工程蓝图（分为：常识教育与专业教育两个阶段）

c、软件系统工程蓝图（分为：通用软件与专用软件两个阶段）

4、 字组分合的基本工具

a、语汇一级的字组细分方法（义项大典）

b、语汇一级的组字重用方法（用例大全）

五、结论

1、可计算的数学模型

2、可选择的数据处理

3、可重用的系统工程

a、语文系统工程（集成--字组细分体系中类似“静力学”的部分）

b、知识系统工程（融智--字组细分体系中类似“运动学”的部分）

c、软件系统工程（共享--字组细分体系中类似“动力学”的部分）

4、可推广的分合方法

a、语汇一级的字组（汉语）或词组（英语）细分方法（单语义项大典）

b、语汇一级的组字（汉语）和组词（英语）重用方法（双语用例大全）

注：多语的情况是上述介绍的单语和双语的通用的模型及其实例的推广。

六、总结

本文通过论述汉语字组细分的计算语言观，重点从基础语言学的角度深入分析了汉语的字和字组的关系，并且，通过比较汉语的字组细分的计算语言观与英语的词组细分的计算语言观，进而，从计算语言学的角度给出了相应的数学公式和计算模型。

字组细分，有狭义和广义之分，狭义的字组细分属于汉语语汇学（基础语言学--主要是指基于汉语字本位的字组细分）和语义词汇学（计算语言学）的交叉学科领域，广义的字组细分涉及文字、语音、语汇、语义、语法、语用、修辞、逻辑、写作的文体和文本的翻译等多学科交叉领域。本文论述狭义的字组细分。

计算模型，是指基于字组细分的计算模型，包括汉语理解的音节总量控制模型（狭义的字组细分与计算模型）与文本总量控制模型（广义的字组细分与计算模型）。本文论述狭义的字组计算。

致谢

感谢北京大学中文系徐通锵教授推荐和邀请本文作者出席“汉语字本位理论专题研讨会”！同时，对徐教授给本文作者在基础语言学方面的指导也深表谢意！

参考文献

1、徐通锵《语言论--语义型语言的结构原理和研究方法》1997东北师范大学出版社

2、徐通锵《基础语言学教程》2000北京大学出版社

3、俞士汶、朱学锋 编《计算语言学文集》2000（第四集）见：<http://icl.pku.edu.cn>

4、邹晓辉《语言及语义信息的统一参照系》2001熵和信息网站 （见：

5、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 》2004第五届汉语词汇语义学研讨会（论文）见：

注释

注1：《语言及语义信息的统一参照系》的修订搞2001年被易绵竹教授收入他主编的（国家重点课题的一个子课题）《计算语言学论文集》之中

注2：2004第五届汉语词汇语义学研讨会（论文集目录）见：

论文摘要（修订稿）

论文题目：论汉语字组的细分（重点分析字和字组的关系）——字组细分方法及其计算模型对汉语字本位理论的意义（作者：邹晓辉）

关键词语：字本位、字组细分、总量控制、人机协作

论文摘要：本文通过论述汉语研究的字组细分方法及总量控制模型，把基础语言学中汉语字本位的观

点推广到计算语言学领域。由于这个推广过程得到了数学和计算机科学等可验证方法的有力论证，因此，不仅在实践方面为后续的语文、知识和软件三大系统工程的顺利开展奠定了语汇一级汉语资源库建设的基础，而且还在理论方面为汉语字本位的观点的推广及其理论体系本身的进一步完善提供了相应的支持或启示。

字组细分方法及其计算模型对汉语字本位理论的意义，即：

1、该方法与模型对汉语字本位理论的操作性，提供实践检验手段；

2、该方法与模型对汉语字本位理论的应用性，提供实践操作工具；

3、该方法与模型对汉语字本位理论的完善性，提供计算分析方法；

4、该方法与模型对汉语字本位理论的普及性，提供系列推广工具。

附注1：据作者所知，本文提及的这项研究，不仅在汉语字本位理论的推广研究领域具有新颖性、创造性和实用性，同时，也未见其他研究者或公开的研究成果。

附注2：该字组细分方法及其计算模型的三个实施例：

a、基于1000个常用字的字组细分和计算试验模型；

b、基于3500个常用字的字组细分和计算试验模型；

c、基于8000个常用字的字组细分和计算试验模型。

上述试验模型的直接现实意义：

一方面，为汉语习得与外汉教学提供协同智能计算工具（印刷版与数字版）；

另方面，也为语文、知识和软件三大系统工程奠定了计算语汇分析的基础。

由此可见，这也间接说明了本研究对汉语字本位理论推广的现实意义。

汉语“字本位”理论专题研讨会论文

字的形式化定义

——试论字本位理论的根基

摘要：本文中心思想是：用“文本体系”或“音节体系”在形式上限定“字的定义”，即：给出“字的形式化定义”。“文本体系”指“文本总量控制模型”。“音节体系”指“音节总量控制模型。作为一种计算机辅助的方式，“文本体系”或“音节体系”，既能以数学形式用于计算机和国际互连网帮助用户学习汉语和机器翻译，同时，也能确立“形式化定义的字”即“作为基本结构（形式）单位的字”在“整个汉语（形式）体系”中“根基”地位。

关键词：字本位、线串型结构（字组）、字的形式化定义

一、绪言

字的定义，属于：基础语言学研究领域。本文的特殊性在于：定义方式的“形式化”，这涉及认知心理学的符号识别方法、计算语言学的自然语言理解方法和人工智能的知识表达方法。本文的重要性在于汉语“字本位”理论的根基是否牢固与字的定义是否能够“消歧（即：消除歧义）”紧密相关，所以，这是当前汉语理论界“立、驳论”双方争议的焦点。本文作者认为：“字的定义”只有“形式化”才能“消歧”，否则，就难以平息“争端”，没有“共识”，哪来（汉语界齐心协力的）“共为”？笔者的研究途径是通过探讨“层面型结构”与“线串型结构”以及“作为两者迭交形式的‘字’的定义”的形式化方法，说明“如何消除‘字的定义’的歧义”以及“什么是字的‘形式化’定义”以及这种方法的“优越性”。本文的限制领域或局限性是：从“形式

化”的角度探讨“汉语的基本结构单位”。其基本假设是：如果汉语的结构单位都是“线串型结构”，而其中的基本结构形式必然位于“线串型结构”的起点或节点，那么“字”就是汉语的基本结构单位，因为，只有“字”正好位于这种“线串型结构”的起点或节点。本研究的贡献在于：不仅从结构形式的角度，为汉语“字本位”理论提供了“字的形式化定义”的方法，可以有效地实现“字的定义”的“消歧”，同时还从概念内容的角度，特别是从内容与形式统一的角度提出了一系列值得进一步探讨的问题。

二、综述

《基础语言学教程》[1]（简称：教程）指出：语言基本结构单位是驾驭语言系统的枢纽。以语言基本结构单位为“纲”，比较汉语和英语等印欧系语言在结构上的异同，揭示不同语言的特点，进而讨论语言结构的基本原理。如：《教程》作者仔细比较研究了汉语、英语、俄语等一些语言的结构单位的异同之后总结出来的假设（即：确定语言基本结构单位的原理），即：语言基本结构单位的三个特点（即：三条标准）：现成性、离散性和语言社团中的心理现实性。

然而，本文作者却发现：此原理，可以进一步研讨。首先，语言，有：形式与内容两个方面，因此，其基本结构单位，也有：形式与内容两个方面，如：符号与概念。其次，特点（标准），也有一个针对性的问题，如：它们是针对：形式或内容，还是两方面都针对？接下来，就自然是：一系列的分析、比较或研究。本文重点进行“形式”方面的研究。

《教程》作者还认为：据此原理可断定：英语等印欧系语言的基本结构单位有两个，即：词和句子，而汉语只有一个，这就是：字。不仅因为，（汉语的）字[或（英语的）词和句子]，是：语言研究应该首先抓住的基本结构单位，而且还因为，以此为基础逐层推进，可进一步弄清楚其它结构单位的性质、特点和它们的各种构造规则。显然。字是汉语“字本位”理论的根基。因为，汉语“字本位”理论，确立了“字”在该理论中的重要地位，具有“根基”的性质。

这个“根基”是否牢固呢？学界同行中有人提出了尖锐的质疑。例如：《计算语言学文集》第 4 集（2000）“汉语语法研究所面临的挑战[2]”一文就正是针对“字的定义”这个汉语“字本位”理论的“根基”提出了十分明确的质疑：

汉语“字本位”理论跟以往的语法理论完全不同。《教程》作者在《语言论》[3]这部专著中一再强调“字”是“汉语句法的基本结构单位”（11 页），“汉语的结构以字为本位，应该以字为基础进行句法结构的研究”（13 页）。质疑者指出：这是“字本位”的核心观点。可是，对“字”这个最核心的概念、使用最频繁的术语却并未给出严格明确的定义，而只是从不同角度做了一些说明。如：“字是形、音、义三位一体的结构单位”（266 页）；“字是汉语结构的枢纽，是语音、语义、词汇、语法的交汇点”（徐通锵 1988a）；“‘字’是汉语对现实进行编码的基本单位”（433 页）；“‘字’是汉语结构的枢纽、结构关联的基点”（433 页）；“字是汉语的基本结构单位，也是最小的结构单位”（434 页）；“我们把字看成汉语句法的基本结构单位”（11 页）；“我们把‘字’定义为：语言中有理据的最小结构单位”（17 页）；等等。这些说明，其含义并不一致（本文作者注：质疑者指出了作为汉语“字本位”理论的“根基”的“字的定义”存在“歧义”），让人难以理解“汉语句法的基本结构单位”的字到底是指什么。“有理据”的含义很不确定。目前哲学界和语言学界对“理据”的理解和看法，因人而异。由此可见，关于字的定义“很缺乏操作性”。这不能不影响人们对字本位理论的认识和理解。

在《语言论》（1997）之后几年出版的《教程》（2001）也说：字义的特点是概括性、民族性、模糊性。似乎并没有针对质疑（本文作者注：“汉语语法研究所面临的挑战”一文在 1998 现代汉语语法学国际学术会议第一次全体大会宣读）做出进一步的应对或论述。

既然如此，怎样才能给“字”这个概念或术语下一个严格而又明确的定义呢？“字的定义”的“操作性”的问题如何才能解决好呢？“字的定义”这个汉语“字本位”理论的“根基”存在的“问题、不足、缺陷或漏洞”能够弥补好吗？

显然，“字的定义”能否“消歧”？这将直接涉及汉语“字本位”理论的“根基”是否牢固的问题。

从汉语界“立、驳论”双方争议的焦点来看，“质疑”是针对汉语“字本位”理论的“根基”而来的。

以下试图科学地回答上述“根基”问题——提供“字的形式化定义方法”，旨在抛砖引玉。

三、方法

首先，确立：大前提。

从方法论的角度，确立本文作者的基本观点，即：对“字的定义”这个汉语“字本位”理论的“根基”问题的解决，不能仅仅建立在“因人而异”的所谓“理解”或“看法”的基础之上。换言之，只有遵循“形式化”的途径，才能把汉语“字本位”理论的“根基”建立在逻辑和数学的坚固基础之上。否则，难免会陷入“公说公有理，婆说婆有理”的非形式化的议论“怪圈”中，即：“靠‘人气’或‘势力’决定‘理论’的优劣”（这显然不是严谨的科学态度和方法！）。

在此，“非形式化的议论”，指：基于内容的议论，它区别于：基于形式的推理和计算，即：“形式化的推理和计算”。众所周知，“非形式化的议论”，往往难以达成“共识”，“形式化的推理和计算”，则容易形成“共为”而不限于达成“共识”。

接着，明确：小前提。

从方法的角度，明确本文作者的基本方法——字的形式化定义方法，即：

（一）前期探索或研究方法（主要是基于文字和表格的准形式化方法）：

a、在“语言及语义信息的统一参照系[4]”一文（2000 年 12 月）中，本文作者指出：汉语的字和英语的词的最大不同是：在“基本笔画、偏旁部首、字、字组”的形式系列中，字以前是“非线性结构”，字以后是“线性结构”而且是多音节，字是单音节且位于前后两种结构的交汇处；在“字母、词素、词、词组”的形式系列中，词的前后都是“线性结构”且词本身既可是单音节也可是双音节甚至还可是多音节。由此而产生其它一系列语言结构形式的不同，其中蕴涵着语言文字的具体构造机理和重用法则。

b、在“协同智能计算语言数据库的设计方法[5]”一文（2002 年 11 月）中，本文作者还以“基本笔画、（不成字、变形字、字中字）三种偏旁部首、字、辞、块、读、句、....”的方式建立了“文本总量控制模型（GTCM）”系列一览表（对应于：相应的线性代数方程组），实施例涉及：0、1、2、3、4、5、6、7、8、9、10、11、12 个基础表，以及它们组成的语言文字数据库。其中，0-12 简称：（自然语言或语言文字的 13 个）进阶层式，体现了语言进化发展不同阶段各个层次的具体结构形式的变化特点和规律。

c、在“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[6] ”一文（2004 年 6 月）中，本文作者指出：1、就古代汉语和现代汉语中与传统一脉相承的语言现象而言，认同“字本位”。2、就现代汉语中吸收西方语言而发生着显著改变的语言现象而言，主张“字组细分”。汉语的混音节“线串型字组”（如：汉语的“辞”“块”，与之对应的是英语的“词”“语”）是单音节“层面型字组”（汉语的“字”）与英语的混音节“词语”之间无歧义连接（同意并列）的纽带（旨在保证“双语”的“义项”形式化“转换”或“对译”）。同时，根据“字组细分”的观点和“拼字音节”的分划方法，把所有的汉语“字组”以单音节的“字”作为“汉语的基本语言（形式）单位”进行计量和排序（1~n），建立“音节总量控制模型（GSCM）”系列一览表（对应于：相应的线性代数方程组），其实施例是：在“字组”（词汇一级），分为：1~n 个系列的“字组细分”一览表[等价于“文本总量控制模型（GTCM）”的第 4~6 进阶的“字组粗分”一览表]。

（二）当前探讨或研究方法（由准形式化发展到纯形式化方法，包括：抽象与直观的方法）：

a、具体策略

一般而言，任何一个字的含义都可能有多个义项，但是，作为一门科学学科的汉语“字本位”理论的“字的定义”，应无歧义。因此，要么从“字”的现有“义项”之中选出一个并确定其含义的“唯一性”（这是基于主体之间的约定方法），要么根据一定参照系给出一个具有唯一性的科学定义（这是基于客体的标准化方法）。本文采用后一种方法。

b、具体途径

汉语“字本位”理论的“字的定义”选定了“基于客体的标准化方法”，因此，确立“参照系”或“标准”的“形式体系”就是唯一可行的途径。只有以此途径确立的“字的形式化定义”才是汉语“字本位”理论的“字”的科学含义。“形式化”是基于“形式体系”而言的。

只有采用“形式化”定义的方法，才能消除所有可能的“歧义”。不给“（有意或无意、主动或被动的）误解”留下任何借口或路径。

c、具体方法

首先，确定：具体的“参照系”，即：“文本总量控制模型（GTCM）”。

表 1 是说明“GTCM”的“0-8”个粗放“进阶层式”一共九个系列一览表的总表。

图 1 是说明代表“GTCM”的“0-8”个粗放“进阶层式”一览表“数码”与对应的“文字”描述或称谓之间一一对应关系的简化示意图。

然后，根据上述参照系的具体形式体系，确定“字”的“形式化”定义。

（1）从两个方向解析“字与字组的关系”的方法，即：从“层面型结构”与“线串型结构”两个方面，或：从表 1 或图 1 中“4”的上下或左右两个方向的解析入手（即：具体的“形式化定义的方法”）。

两个方向：

一个方向是逆向，指：在表 1 或图 1 的“进阶层式”中，“4”与“0、1、2、3、4”关系；

一个方向是正向，指：在表 1 或图 1 的“进阶层式”中，“4”与“4、5、6、7、8”关系。

什么是“层面型结构”？

“层面型结构”，指：位于“0、1、2、3、4”诸“进阶层式”的成员，因其结构形式在计算机分析过程中呈现出“层面特征”而得名。

什么是“线串型结构”？

“线串型结构”，指：位于“4、5、6、7、8”诸“进阶层式”的成员，因其结构形式在计算机分析过程中表现出“线串特征”而得名。

（2）从“层面型结构”与“线串型结构”的“迭交”之处，解析“字与字组的关系”的方法。

“层面型结构”与“线串型结构”如何“迭交（即：交叉重叠）”？

“迭交”，指：在表 1 或图 1 的“进阶层式”中，位于“0、1、2、3、4”与“4、5、6、7、8”的“迭交”之处的“4”这个唯一的“进阶层式”成员（即：“字”的结构形式）。

通俗的讲就是：

当位于“0、1、2、3、4”终点的时候，“字”作为“层面型结构”（涉及：字内的文字符号之间的组合）的构造形式，表现为：“文字”这一结构形式；当位于“4、5、6、7、8” 起点的时候，“字”作为“线串型结构”（涉及：字外的字组符号之间的组合）的构造形式，表现为：“语言”的结构形式，而且是：“基本结构（形式）”。

（3）从“层面型结构”与“线串型结构”迭交的“条件”来看，“字”这个特殊的结构形式可以作为汉语的其它“结构形式”的计量“单位”，解析“字与字组的关系”的方法。

“层面型结构”与“线串型结构”迭交的“条件”是什么？

迭交的“条件”是：在位于“0、1、2、3、4”的“层面型结构”与“4、5、6、7、8”的“线串型结构”之间，同时存在“4”即“字”这个位于“迭交之处”的“进阶层式”。

进一步，因为“4”即“字”这一“进阶层式”的“结构形式”在形式上正好是构成“5、6”即“辞、块”的“基本形式”，具有“可重复、可测量、可计算”的形式特征，所以，就可在语

汇一级把较为粗放的“参照系”[即：“文本总量控制模型（GTCM）”]发展成为精细的“参照系”，即：“音节总量控制模型（GSCM）”，从而在“4、5、6”即“字、辞、块”的范围，实现：基于“字”这一“基本结构（形式）单位”的“字组细分”。

四、结果

字的形式化定义：

定义 1

汉语“字本位”理论所述的“字”，特指：位于“文本总量控制模型（GTCM）”第“4” 这个特定序位的“进阶层式”的所有单个的“汉语结构形式”。

表 1 是说明“GTCM”的“0-8”个粗放“进阶层式”一共九个系列一览表的总表。

编号	进阶	汉语	拼音	英语
1	0	基本笔画	字母表	26 个字母
2	1	不成字偏旁部首		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	4	单音节的“字”（基本结构形式单位）	单音节	混音节单词
6	5	复音节的“辞”（字组）离心与向心	多音节	多音节词组
7	6	多音节的“块”（字组）含两种成份	多音节	多音节短语
8	7	标逗号的“读”（表示：语气上的停顿）	逗号	逗号
9	8	标句号的“句”（表示：语义上的停顿）	句号	句号

表 1

图 1 是说明代表 GTCM 的“0-8”个粗放“进阶层式”一览表的具体“数码”与对应的“文字”描述或称谓之间一一对应关系的简化示意图。

0	1	2	3	4	5	6	7	8
笔画	不成字部首	变形字部首	字中字部首	字	辞	块	读	句

图 1

定义 2

汉语“字本位”理论所述的“字”，特指：位于“音节总量控制模型（GSCM）”第“1”这个特定序位的“进阶层式”的所有单个的“汉语结构形式”。

公式 1 是抽象表示“GSCM”的数学形式，说明“1-n”个精细“进阶层式”的“n”（多）个系列一览表及其总表，都可以表示为：线性代数方程（组）。

$$\sum a_{ij}x_j = b_i \quad \text{等价于：一组“字符多项式”[7]，即：} A = \{\sum n_i x_i\}^5$$

公式 1

图 2 是说明代表 GSCM“1-n”个精细“进阶层式”一览表的具体“数码”与对应的“文字”描述或称谓之间一一对应关系的简化示意图。

1	2	3	……..	n
一字组	二字组	三字组	……..	多字组

图 2

上述定义 1 和定义 2 是等价的。由此定义的“字”的集合，不仅包括了 Unicode 中的所有“汉字”，

[5] 或等价于：一系列“字组数据表”，即：与图 2 所述的系列“字组”表对应。

而且，还包括了所有将归入 Unicode 中的“字”或“汉字”，也包括“FONTS”中用于“显示”、“打印”或“输出”的“字”或“汉字”。它们是构成“字组”或“线串型结构”的基本结构（形式）单位。

据此生成的“字的定义”的优越性表现在哪里？

据此生成的“字的定义”的优越性表现在它的“无歧义性”。

例如：位于“文本总量控制模型（GTCM）”第 4“进阶层式”的“字”，一方面，可视为：第“0、1、2、3、4”进阶层式的高端或终点，另一方面，又可视为：第“4、5、6、7、8、...”进阶层式的低端或起点。因为，位于第“0、1、2、3、4”进阶层式的“基本笔画、（不成字、变形字、字中字）三种偏旁部首、字”，充其量都只是限于“方格”或“方块”之中的“层面型结构”的某个页面或层面；而位于第“4、5、6、7、8、...”等一系列粗放“进阶层式”的“字、辞、语、读、句、....”或位于“音节总量控制模型（GSCM）”第“1，2，3，…” 等精细“进阶层式”的“一字组、二字组、三字组、…”则是由上述“方格”或“方块”组成的“线串型结构”的某一占据“1、2、3、.....”个方格“线性字串”。其中，只有“字”同时位于“文本总量控制模型（GTCM）”（第 4“进阶层式”这个特定序位）与“音节总量控制模型（GSCM）”（第 1 个一览表）“交集或并集”的“单音节”一览表之中。显而易见，“字”既属于“层面型结构”的高端或终点，又属于“线串型结构”的低端或起点，而且，两个端点是“迭交”在一起的。

定义 3

从基础语言学与计算语言学结合的观点来看，也可以说：“字”，作为汉语的基本结构（形式）单位，特指位于“层面型结构”与“线串型结构”的“迭交之处”的那类汉语结构形式，其特征在于：1、单音节，2、方块形，3、多义项（其中，各个“义项”的“等价”形式，就是一系列与各个“义项”一一对应的“字组”，详细内容将在与本文的“姊妹篇”——“字组的划分方法”和与之一同构成“三部曲”的“字与字组的关系”两篇文章中涉及“义项解释形式化”或“字组化”的“义项本位”部分专门论述）。

图 3 说明“GTCM”与“GSCM”的两个“参照系”同时给出的“字的形式化定义”示意图。

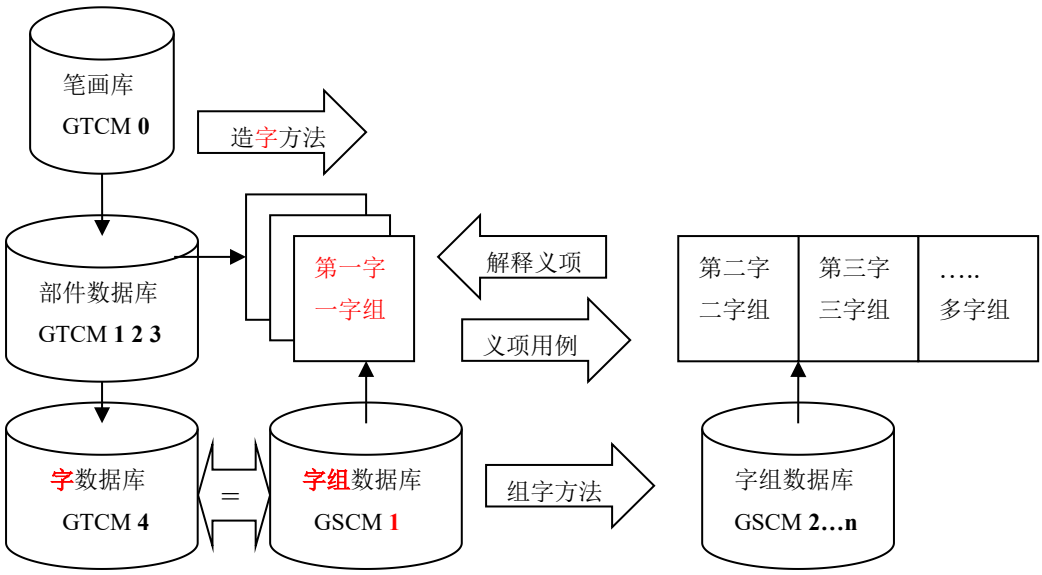


图 3

上述定义 1、定义 2 和定义 3 是等价的。它们表面上展示“字与字组的关系”，实际上背后都有相应的数据库在支持。

五、结论

综上所述，本文给出“字的形式化定义”的前提和结论如下：

1、大前提

任何一种语言结构（形式）单位，均可以“文本总量控制模型（GTCM）”与“音节总量控制模型（GSCM）”的方式“（间接）形式化”。

2、小前提

以“文本总量控制模型（GTCM）”与“音节总量控制模型（GSCM）”（间接）形式化的各个语言结构单位，在“GTCM”与“GSCM”这两个“参照系”中，均有其特定的“序位”。

3、结论

“字”作为一种特定的语言结构（形式）单位，在“文本总量控制模型（GTCM）”与“音节总量控制模型（GSCM）”中，也有其特定的“序位”。

由此可见，本文给出“字的形式化定义”是唯一的、无歧义的。

六、讨论

笔者认为：虽然从形式结构方面看，汉语的基本结构（形式）单位只能是字，这将成为一个显而易见的事实。但是，如果从概念内容方面看，问题并没有这么简单。如果从内容与形式统一的角度看，那么，至少还有以下一系列问题或相应的情况值得进一步关注、探讨或交流。

1、如果仅就古代汉语和现代汉语一脉相承的语言现象而言（例如：成语），那么，“字本位”显然是成立的。

2、如果就现代汉语中吸收了西方语言的东西而发生显著改变的语言现象而言，那么，“本位”问题会比较复杂。其理由如下：

a、首先，“字本位”、“词本位”、“短语本位”、“小句本位”等虽然旨在讨论汉语的基本结构（形式）单位，但是，却都夹杂了“非形式化”的东西（语言形式承载的思想内容）。

b、其次，如果仅从形式方面来区分汉语的基本结构（形式）单位，那么，问题会非常清楚。例如：本研究对“字”和“字组”的定义（其中“字组”部分见本文的姊妹篇“字组的划分方法”），就是这样的，都是采用“（间接）形式化”方法或途径。

c、再次，必须指出：上述各“本位说”的区别主要在于各自形式的不同，如一旦涉及内容问题立即变得复杂起来。例如：当“字”、“词”、“短语”、“小句”等表示“相同的概念”或指称“同一个物象”的时候，立即产生“形式歧义”（而非“内容歧义”）。

d、最后，必须指出：就表达“概念”或“对象”而言，“字”、“词”、“语”的区别主要在于“内容”方面（见“字与字组的关系”）而非“形式”方面。

3、从整体上看，无论是“字本位”还是“词本位”或是“短语本位”乃至“小句本位”，都恰似“盲人摸象”一样，仅仅摸到了（汉语这个）“大象”的一个（非形式化的）部分。

4、尽管如此，这仍是非常了不起的！因为（汉语这个）“大象”的确太大，致使任何个人的经历或阅历要想统观全局且一览无余都难以想象。

总而言之，如果单从结构形式方面看，那么，“字本位”显然有其特殊的优势。

众所周知，凡有多个结构（形式）单位存在，就必定存在一个基本结构（形式）单位。由于从外语借用或导入的“词”，对汉语来说，可是“一字组（即：单个的字）、二字组、三字组、...多字组”中的任何一个，“短语、小句”的道理也如此，只有“字”位于最基本的位置。所以仅就形式而论，汉语的基本结构（形式）单位，非“字”莫属。

这样看来，“字的形式化定义”的确可解决“字本位”理论的“根基”的定性问题。

接下来，就该探讨“字组的划分方法”以及“字组”的计算或定量分析问题了。再进一步，才便于较为深入地研究“字与字组的关系”这一既涉及结构又涉及程序的更加复杂的问题。

致谢

北京大学中文系基础语言学教研室徐通锵教授与中国人民解放军洛阳外国语学院计算语言学教研室易绵竹教授阅读了本文的初稿，前者从基础语言学专家读者的角度给予了作者珍贵的肯定，后者从计算语言学专家读者的角度给予了作者宝贵的鼓励，在此作者向他们表示真诚的谢意！

同时，还要感谢作者的母亲和妻子给予的关怀和帮助！

没有上述各位的关心和帮助，此文难以在如此短的时间内修订完稿并与汉语“字本位”理论专题研讨会的各位专家见面（指：论文公开）。希望它能起到“抛砖引玉”的作用！

最后，还要感谢汉语“字本位”理论专题研讨会组委会提供这次机会！

参考文献

1、徐通锵《基础语言学教程》2001 年 2 月北京大学出版社，19-36 页，178-237 页[M]

2、北京大学计算语言学研究所《计算语言学文集》第 4 集“汉语语法研究所面临的挑战（98 现代汉语语法学国际学术会议第一次全体大会宣读）”2000 年 12 月，1-19 页[C]

3、徐通锵《语言论--语义型语言的结构原理和研究方法》1997年10月东北师范大学出版社，295-442页[M]

4、邹晓辉“语言及语义信息的统一参照系（2000 年 12 月）光明网论文发表交流中心转载

[EB]

5、邹晓辉“协同智能计算语言数据库的设计方法（2002 年 11 月）--支持语言文字系统工程与全球语言定位系统的一个实施例”光明网论文发表交流中心转载

[EB]

6、《第五届(国际)汉语词汇语义学研讨会论文集》“义项语汇典例（SVDE）的总量控制模型（2004 年 6 月） --人机协作对采用汉语注释的语义词汇典例进行计量分析”光明网论文发表交流中心转载

[EB] [EB]

7、张学文《组成论》附录 1“字符多项式与表格数学”中国科学技术大学出版社 2003 年 12，44-56 页，246-252 页[M]

[EB]

汉语“字本位”理论专题研讨会论文

字组的划分方法

——试论字本位理论的功用

摘要：本文的中心思想是：用“音节总量控制模型”限定“字组的定义或划分”，即：给出“字组的划分方法”。本文是“字的形式化定义”一文的“姊妹篇”。首先，在形式体系中，采用“字”作为基本结构（形式）单位，对“字组”进行划分，即：“字组细分”。进而从语言学家、普通用户、（领域）专家那里获取关于字组划分的信息或知识。因此，借助计算机和国际互连网，根据“音节总量控制模型”，所有用户可在任何时候或任何地方查询或重用“字组”。这也就是汉语“字本位”理论功用的一种典型体现。

关键词：字本位、线串型结构、字组的划分方法

一、绪言

本文作为“字的形式化定义”的姊妹篇，进一步给出“字组的形式化定义”或“字组划分的标准化方法”，涉及：基础语言学与计算语言学交叉的研究领域。其特殊性在于“字组定义形式化”或“字组划分标准化”，不仅涉及符号识别、自然语言理解和知识表达，而且还涉及代数多

项式和字符多项式的应用。其重要性在于：在语汇的形式处理方面，体现了汉语“字本位”理论的功用。由于“字组定义存在歧义”或“字组划分缺乏标准”是当前汉语理论界“立、驳论”双方争议的另一个焦点，所以，能否解决此问题是能否消除歧义、平息争端、达成共识、形成（汉语界齐心协力的）共为的关键。其研究途径是探讨“线串型结构”论述“字组定义形式化”或“字组划分标准化”及其优越性。其局限性在于仅从形式化与标准化的角度探讨字组的定义与划分。其基本假设是：把“字”视为“线串型结构”的“节点”（含“起点”），于是，字既可被视为组配字组的基本结构（形式）单位，也可被视为一种特殊的（次广义的）字组，即：一字组[6]。其贡献在于：为汉语“字本位”理论提供“字组的划分方法”，并在排除“字组定义”歧义的同时，确立“字组划分”标准，以及“字组”查询或重用的计算方法及公式。

二、综述

根据汉语“字本位”理论[1]可归纳出以下关于“字组”的观点、原理和方法：

就字组的含义而言，把“字的组合”称为字组，并视为“从字到句的过渡性结构单位”。

就字组的分类而言，一方面，把字组区分（即：细分）为：二字组（作为重点考察对象）、三字组、四字组、...多字组；另一方面，又把字组区分（即：粗分）为：表达概念（辞）与表达结构（块）两种基本类型（属于：语汇一级的狭义字组，也是通常意义的字组，一般不需特别说明）。

就字组的构成而言，提出以核心字（即：在字组中位于核心地位的那个字）为向心与离心的组字法。

就字组的例证而言，指出《现代汉语词典》与《倒叙现代汉语词典》实际上就是以核心字为基础而编制的两本研究字组结构和字义关系的重要工具书。

《计算语言学文集》第4集“汉语语法研究所面临的挑战[2]”一文对汉语“字本位”理论的质疑，不仅涉及“字”（注1），还涉及其他一些重要的概念或术语，如，指出：在《语言论》[3]这一“字本位”专著中“字组”（353页）、“核心字”（365页）等，也都没有进行严格而又明确的定义（注2）。因此，影响了人们对汉语“字本位”理论的认识和理解。

可见，字组定义或划分存在歧义是当前汉语界“立、驳论”双方争议的另一个焦点。如不能消除这一歧义，就难以平息争端，而不平息争端，就难以达成共识，这样要形成汉语界齐心协力共为的局面就有困难。

因此，能否根据语言事实说清楚“字组定义形式化”或“字组划分标准化”就是很重要的。本文基于工程和应用两方面实践而提出以下建设性方法或方案。

三、方法

1、方法概述

汉语“字本位”理论所述的“字组”包括“辞”与“块”，可视为狭义的字组，在“文本总量控制模型（GTCM）”（注：本文仅讨论其汉语部分）中，位于“第5和6进阶层式[*1]”。广义的字组，由“GTCM的第5和6进阶层式”向两端延伸而来，如：逆向延伸其低端至“GTCM的第4进阶层式”便形成限定在语汇一级的次广义的字组，涉及：“字、辞、块”；“音节总量控制模型（GSCM）”就是以位于“GTCM的第4进阶层式”即“字”为“基本结构（形式）单位”对“GTCM的第5和6进阶层式的狭义字组”实施“形式上的细分”而构成的。在此基础之上，就可深入一步对汉语语汇（即：次广义的字组）进行“字组细分”——涉及内容上的理解[通过“语言信息标注、常识信息标注、（学科或领域）知识信息标注”而实现。这将有利于：“海量的人类基本知识形式化表现”（如：CYC常识知识库、wordnet概念词典、encyclopedia百科全书）和“专家系统”（ES）的快速构建和及时改进或优化，同时，也将有利于：“计算机辅助（CA）汉语教学”和“机器翻译”以及“其

[6] 实际上是解释语言中的字（区别于对象或符号语言的字）。

它中文信息处理项目”在内容和形式的整合方面实施各种进一步的“字组划分”（具体作法在“工程融智学与应用融智学双向互动及优化实践（一种新颖的产生式教学——针对双语处理及其一体化管理——针对知识处理）”的系列专论中介绍）]。

2、方法详述

（一）总体策略：

通过“字组定义形式化或字组划分标准化”为下一步实现“义项解释字组化”奠定坚实基础。

可选方法或途径：

a、着重研究语言事实**外在现象**方面的归纳方法（常规途径之一），

b、着重研究语言事实**内在机理**方面的演绎方法（常规途径之二），

c、强调**内外统一**的比较与穷举乃至对归纳和演绎的整合方法（常规途径之三），

d、强调借助计算机的运算处理速度和海量记忆存储能力以及通信网络的高效传输能力乃至便于输入输出反馈的人机友好协作的交互界面，特别是对语言事实（含：外在现象与内在机理）进行科学的归纳、演绎、比较、穷举。本研究采用“d”方法或途径，并以形式化和标准化的工具或模型作为具体的实现手段。

（二）前期探索：

由于“协同智能计算语言数据库的设计方法[4]”一文（2002年11月）在“文本总量控制模型（GTCM）第5、6进阶层式”，仅仅涉及“辞、语”的**字组粗分**方式，所以，“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[5]”一文（2004年6月）进一步提出了“音节总量控制模型（GSCM）”的**字组细分**和**拼字音节分划**的标准化方法，即：把（次广义的）**字组**以**单音节的“字”**作为**汉语的“基本结构（形式）单位”**进行**计量**，把字组数据库各个表及其**行**和**列**均按自然数id自动编号对每一个格实现自动测序定位，通过“三级标注[*2]”实现各种有针对性的序位变换或重组排序以满足用户（标准化、个性化、标准化与个性化两方面结合的各种）需求。

“**音节**总量控制模型（GSCM）”多个系列（次广义的）字组**细分**一览表在总量上等价于“**文本**总量控制模型（GTCM）第4-6三个进阶层式”的（次广义的）字组**粗分**一览表。

中文信息处理领域一直以来都缺乏一套既便于前台直接显示**自然语言**文字又便于后台直接计算**自然数**的完整数学方法或工具（好比是一个可以称量语言和知识的超级“天平”）。虽然与GSCM细分字组一览表id集合等价的“代数多项式”或“线性方程组”是实现总量控制的关键，但与GSCM细分字组一览表等价的“字符多项式[6]”却更适合上述条件。

（三）方法步骤：

本文将进一步对汉语的“线串型结构”做深入细致的分析。

a、确定字组**划分**标准——参照系，即：

（1）字组粗分参照系：“文本总量控制模型（GTCM）”——“字组形式化定义”的依据；

（2）字组细分参照系：“音节总量控制模型（GSCM）”——“字组标准化划分”的依据。

b、根据标准划分字组——粗分与细分

（1）（广义）字组划分

首先，以“文本总量控制模型（GTCM）”为“参照系”，以“GTCM第4进阶层式”的“字”为“界”或“中介”，划分“层面型结构（如：GTCM第0-4进阶层式）”与“线串型结构（如：GTCM第4-8进阶层式）”。分别从逆向（即：由第4向第0进阶层式）与正向（即：由第4向第8进阶层式）两个方向，对汉语结构（形式）实施具体的划分。

“层面型结构”，涉及：“笔画字”和“缺损字、变形字、字中字”[均为：偏旁部首，属于：文字学和造字法的研究领域，可视为：不发音的（特殊的）无音（广义的）字组]以及“（标准）字”。“线串型结构”，涉及：“（标准）字”（单音节）和“辞、块”[“双、多”音节，属于：要发音的（通常的）有音（狭义）字组]以及“读、句、…”。其中，“字、辞、块”只是粗分（次

广义的）字组。

图 1 是“（广义）字组划分”示意图。

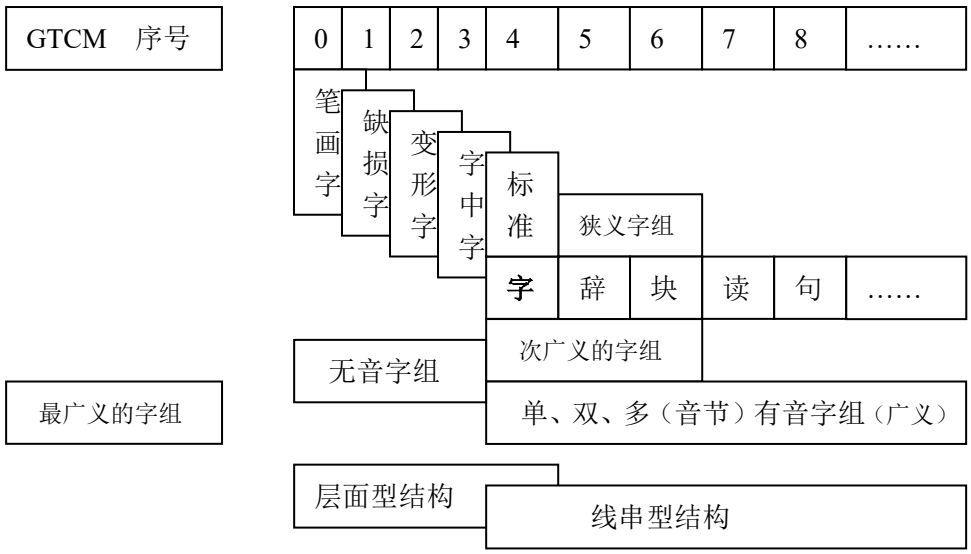


图1 字组粗分示意图

在图1中， GTCM序号表示进阶层式。

（2）细分（次广义的）字组

图 2 是“（次广义的）字组细分”示意图。

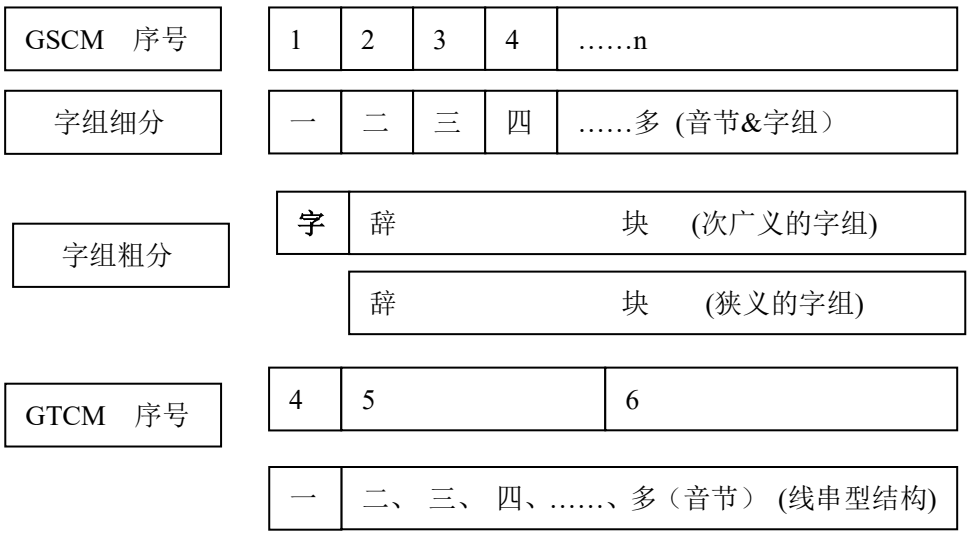


图2 字组细分示意图

在图2中，GSCM序号1，2，3，4，...，n对应地表示“一、二、三、四、...、多（次广义的）字组”一览表的系列序号。以“音节总量控制模型（GSCM）”为“参照系”，以“GSCM序号为1的一览表中的‘字’为‘起点’”（准确地说是包含“起点”的“节点”或“划分尺度”）沿正侧或正向（即：由1向2，3，4，...，n的方向）逐一对“线串型结构”进行划分或细分，即：分别按照“GSCM序号（1，2，3，...，n）的系列一览表，存取“一（次广义的[7]）、二、三、四、...、

[7] 字组中的字（解释语言）与独立的字（符号语言），实质上分属释义元语和对象语言两个学科研究领域。对此，

多字组”。

（3）形式化与标准化

为了进一步深入细致地探讨“（次广义的）字组”的“分与合”的机理、机制或法则，首先，“形式化”细分字组，然后，“标准化”细分字组，即：构造“基于细分字组的**电子表格**或**展示工具**”（即：细分字组**各个系列**的一览表）；建立“基于（次广义的）**字组细分**的数学通式或计算模型”（即：$\sum a_{ij}x_j=b_i$ 系列一览表的id）。

其核心步骤的简化转述：建立“系列一览表的id（自然数）与字符（自然语言）一一对应的“（次广义的）字组细分一览表”。从“简化表述”或“直观实用”的角度考虑，可把该展示工具（电子表格）与该计算模型（数学公式）之间的对应关系，转述为一系列“字符多项式”的形式，即：$\mathbf{A}=\{\sum n_i x_i\}$。

四、结果

1、（狭义）字组定义的形式化

汉语“字本位”理论所述的“字组”，即：“辞、块”，是：狭义字组，其形式化定义，特指：位于“文本总量控制模型（GTCM）第5、6进阶层式”这两个特定序位（即：GTCM序号5、6）的类——汉语结构（形式）。

2、（狭义）字组划分的标准化

在“文本总量控制模型（GTCM）”中，（狭义）字组粗分为：辞、块[8]。

在“音节总量控制模型（GSCM）”中，（狭义）字组细分为：二字组、三字组、四字组、...、多字组。

3、（次广义的）字组的形式化与标准化

不仅GSCM序号“1，2，3，4，...，n ”与“一、二、三、四、...、多（次广义的）字组的系列一览表”之间是一一对应的关系，而且，其中的各个系列一览表的id在$\sum a_{ij}x_j=b_i$中的序位也都是唯一的。

（1）（次广义的）字组细分形式化，即：表格化。便于字组表示和标注的自动化。

（2）（次广义的）字组细分标准化，即：公式化。便于字组计算和统计的数学化。

4、（次广义的）字组的表示与计算

（1）用“id”（自然数）与“（次广义的）字组”（自然语言）“同义并列”且“一一对应”的方式，构造“细分（次广义的）字组”各个系列的一览表。

（2）其中，表格的“id”（自然数）部分与数学模型（即$\sum a_{ij}x_j=b_i$）的“特式”或“特例”（即：具体的算法和数据结构）等价。这是实现“（次广义的）字组”间接计算（批处理）的基础。

（3）其中，表格的“字组”（自然语言）部分与展示界面（如：以“重用、组合、轮排”等方式存取或调用）的标准化或个性化（即：由具体的应用软件体现）等价。这是实现“（次广义的）字组”直接表达（自动化或计算机辅助）的基础。

（4）用“字符多项式”表达（2）计算模型（代数公式）与（3）展示工具（电子表格）之间的对应关系，便于寻找具体的分布函数或求解（$\sum a_{ij}x_j=b_i$）的“特式”或“特例”。

5、用通俗的话转述（狭义）字组的形式化定义

从理论语言学与计算语言学结合的观点来看，“（狭义）字组”，作为汉语结构（形式），特指：由“字”按照一定的规则组配的“线串型结构”，其特征在于：1、混音节（含：双音节、三音节、四音节、…、多音节），2、线串形（如：由方块字组配的字符串，汉语拼音字符串与之

笔者在作为《字本位与中文信息处理的基础》的融智学导论的语言学基础研究部分有专门的介绍或论述。

[8] 在《字本位与中文信息处理的基础》的语言学基础研究部分笔者增加了“链”，从而明确了“辞——由实字构成的字组、链——由虚字构成的字组、块——由实字和虚字共同构成的字组”三个属于狭义的字组的基本范畴。

同义并列），3、义项的内容与形式之间呈“反变”关系，即（狭义）字组，随着结构（形式）的增长，其义项（内容）反而减少。也就是说具有义项发散性的字的多义项，将随着（狭义）字组的字数的增加而使组配的（狭义）字组的义项趋于收敛，即：外延大内涵小。“（狭义）字组”与“字”的本质区别在于：“（狭义）字组”可是多种结构形式，如：二字、三字、...、多字，而“字”只有一种结构形式（即：一字）——因此它作为“线串型结构（形式）”的基本计量单位具有唯一性。

五、结论

综上所述，就“字、辞、块、读、句”而言，不仅作为汉语的基本结构（形式）单位的只能是形式化定义的“字”，而且，汉语在语汇一级的其它结构（形式），如“辞、块”，或“二字组、三字组、四字组、...多字组”乃至扩展到汉语在句法一级的其它结构（形式），如“读、句”，或“二字句、三字句、四字句、...多字句”，都可通过“字”作为“基本结构（形式）单位”进行层层分解或组合。

结论：汉语的其它结构形式，如“辞、块、读、句”等，没有一个比“字”更适合做“汉语的基本结构（形式）单位”。

（狭义）字组的形式化定义或划分，涉及计算语言学中文信息处理领域的所谓“汉语分词”难题[因为汉语中文没有与英语英文之类拼音语言文字一一对应的“词”这个结构（形式），所以，对中文信息处理而言，“词的切分”是学界公认的难题！]，对“辞、块”这一“粗分的（狭义）字组”很难实施形式化处理（因为存在虚实兼用或一个字组多用的情形）就是一个典型的实例。

可是，一旦实现“粗分（狭义）字组”与“细分（狭义）字组”的转换，情况就会大不一样。因为，“字组细分”之后，容易实施形式化和标准化处理，从而，为进一步实施个性化重用奠定了计算机辅助乃至自动化应用的坚实基础。这也就是基于汉语“字本位”理论的“（狭义）字组细分”的“形式化”与“标准化”的好处或优越性。具体应用举例如下：

1、在中文信息处理领域，可这样解决“（通常所说的）汉语分词”难题，即（间接形式化）：

a、变“字组粗分”为“字组细分”，实现“形式化”和“标准化”；

b、变“（随时可落入‘语义泥潭’的所谓）汉语分词”为“（完全形式化和标准化的）汉语字组细分”，以便于计算机辅助和自动化标注：语言信息、常识信息、（学科或领域）知识信息。

c、进一步可构成：基于汉语（狭义）字组细分和“三级标注”的“静态字组”数据仓库。

d、在“静态字组”数据仓库的使用过程中，再进一步发展出相应的“动态字组”数据库，其中，重用部分，完全自动化；创新部分，全面实现“计算机辅助（CA）”。

2、在汉语教学领域，可这样优化教学过程及效果，即：

a、借助[基于“音节总量控制模型（GSCM）”的]“细分字组”数据（仓）库集大成的“语言信息、常识信息、（学科或领域）知识信息”，改进或优化“计算机辅助（CA）教学”，如：改进或优化（作为母语的）“汉语教学”与（作为外语的）“汉外教学”的过程及结果。

一方面，充分利用计算机批处理或自动化的高效率支持“计算机辅助（CA）教学”；另一方面，充分借鉴或融合各类智能系统或领域专家们标注或精选的“语言信息、常识信息、（学科或领域）知识信息”，在教学实践中不断改进或优化“二字组、三字组、四字组、...多字组”（乃至“二字句、三字句、四字句、...多字句”）等形式化或标准化的“汉语结构（形式）”——“动态字组”数据库与静态字组”数据仓库，不断完善“计算机辅助（CA）标注（其中：个性化的标注与教学活动息息相关）”、“海量知识的形式化表示”（CYC）和“专家系统”（ES）。

通过上述两方面（如：“动态字组”数据库与“静态字组”数据仓库，与“CA”“CYC”“ES”等各种界面及其调用程序或智能代理）的优化互动，既可帮助用户改进或优化“教学或训练”的过程及效果[如：熟练掌握汉语“组字成语、组字成句、组语成句”的方法（其中包含汉语自身的

一系列规律或法则，如：语法），达到“以简驾繁”的目的或效果]，又可促使系统本身（如“动态字组”数据库与静态字组”数据仓库及其后台软件程序和前台交互界面）不断地改进或优化。

b、通过“人人、人机、机机、机人”之间的高度协作和优势互补[9]，把前述“以简驾繁”的过程及效果提升到足够“目标用户群”共享的程度，从而，显著地提高计算机辅助（CA）教学的智能化水平。

有了字和字组的形式化定义，作为汉语“字本位”理论根基的“字”的含义和汉语“字本位”理论应用的“字组”的含义（涉及：字组的定义或划分）也就都不再有歧义。汉语理论界也没有必要就形式“本位”问题而继续争论。从此，不仅汉语“字本位”理论的“根基”（在形式方面）牢固了，而且，汉语“字本位”理论的功用（在字组划分方面）也明确了。

对字和字组的形式化定义的探讨，不仅涉及“字组”的合成，还涉及“字组”的分解（如：字组的粗分和细分）。“字组”以“字”为基本结构（形式）单位，进行的划分是可计量的。

如果“辞、块”可视为“粗分（狭义）字组”的类型，那么，“二字组、三字组、四字组...多字组”则可视为“细分（狭义）字组”的基本类型。

如果“字、辞、块”可视为“粗分（次广义）字组”的类型，那么，“一字组（即：单个的字）、二字组、三字组、四字组...多字组”则可视为“细分（次广义）字组”的基本类型。

如果“字、辞、块、读、句”可视为“粗分（广义）字组”的类型，那么，其中，次广义的字组（如：字、辞、块），属于语汇的范畴；广义字组（如：读、句），则属于句法的范畴。

由于从（狭义）“字组”到“字”的分解过程的“进阶层式化”，特别是语汇一级“（次广义的）字组”结构形式的数量极限“可计算”，因此，确保在计算机辅助（CA）的条件下可实现总量控制。

随着汉语“字本位”理论的定性问题（字和字组定义形式化）和定量问题（字组划分标准化）的解决，“语言文字系统工程”和与之密切相关的“学科知识系统工程”也将提上议事日程。如：用于解释“字”的“义项”的“二字组（几十万）、三字组、四字组、...、多字组”细分一览表（这一形式化和标准化的语言结构形式体系，作者虽已初步实现但还有待在进一步的实验、中试和商品化三个层级的科技经济推广过程中逐步改进或优化！）如与“产生式教学”和“一体化管理”实现优化互动，那么，“语言文字信息、通用常识信息、学科或领域的专用知识信息”的协同标注，也就将随之而显著改观，即：由“以往只有少数专家与其弟子或助手参与的作坊式知识加工”向“将来还有广大用户和各个领域专家与其弟子或助手协同参与的网络式知识加工”转变。

上述蓝图如能一步一步地实现或推进，那么，改进或优化之后的整个汉语“字本位”理论的大厦，就不仅仅是建立在逻辑和数学抽象演绎的基础上，特别是：通过计算机科学技术的形式化处理，在汉语教学、汉外教学与中文信息处理的诸多实际应用中，都可发挥出它应有的功用。如：虽然汉语拼音只有400多个音节，加上声调的变化，总共也不过1000多个，但与同一“字音”并列的“音字”却并非都是一个，严格讲汉语“音字”的总数至少有几万个，因此，如果我们不采用“相对完全归纳”的策略和借助“进阶层式化”的字组划分方法就消除不了音形结合的汉语歧义。如果消除汉语歧义的形式方面，只涉及常用汉字和次常用汉字共3500个，通用汉字8000多个，Unicode列举的汉字2万多个，那么，在内容（字的义项）方面，却要涉及可用自然语言描述并且公开公知的“语言文字信息、通用常识信息、学科或领域的专用知识信息”的巨大数量（虽然这对一个人、一个单位、一个地方来说，是一系列的天文数字，但是，对 一个跨国公司、一个国家、一个民族而言，还是能通过采用“产生式教学[10]”和“一体化管理[11]”的过程来与“语言文字系统工程”和“学科知识系统工程”之间形成优化互动从而大大地简化前述的“特式”或“特例”的具体分析和计算。尤其是在计算机和互连网普及的时代，如能充分利用本文的方法，那么，原先

---

[9] 即《融智学》的16字方针：（“人人、人机、机机、机人”的）合理分工、优势互补，高度协作、优化互动。

[10]另见“产生式教学法”（作者：邹晓辉），涉及的试验应用小组，如：某计算机信息学院知识管理研究生，又如：某外国语学院英语教学本科生，吉林大学珠海学院英语教研室教师，温州职业技术学院研究所的应用课题，…

[11] （教育领域的）产学研（人）用（机）算必然涉及知识的一体化管理。

的天文数字也就可转化为指日可待的实际工程项目的具体进度指标（涉及化无限为有限的途经）。

六、讨论

借助计算机及其互连网，既利于人们发现“创新字组”，也利于人们重用“现有字组”。从而，借助计算机辅助（CA），显著地提高“研究、教学、生产、应用、计算”的效率及效能。

例1，就汉语“字组”的两种组合方式[7]而论，还可进一步探讨或细化。

直接组合，以语序为主要手段，把两个或两个以上的字组直接组合成一个较大的字组。如：组字成语、组语成句、组字成句。涉及“联合、修饰、陈述、动宾、补充”等组合关系。

关联组合，以起关联作用的虚字（词）为主要手段，把两个或两个以上的字组连接起来组成一个较大的字组。如：组字成语、组字成句、组语成句。涉及“并列、承接、递进、选择、转折、因果、假设、条件”等组合关系。

例2，就汉语“字组”的强制搭配[8]而论，对固定搭配和特定搭配（含：控制性搭配与呼应性搭配）的语义关系类别及其性质，还可进一步做计算机辅助（CA）研究或探讨。

例3，就英汉语汇对比[9]而论，广义语汇学（包括：语义学、辞源学、字典学、修辞学）与狭义语汇学（“字、辞、块”的性质、构成、意义及发展、语汇的构成及发展等规律），可进一步做计算机辅助（CA）研究或探讨。

例4，就在计算机上实行汉语的双轨制[10]而论，也可进一步研究或探讨汉语字组及拼音形式与英语词语及音标形式之间如何建立一一对应关系的问题。这是一个典型的人机协同研究或实验。

例5，就语言理论而论，对语音、语义、词汇、语法“四大要素”与语义、语法、语用“三个平面”乃至语音、语义、语法“三个方面”、“两层结构”（即：表层结构与深层结构）等各种语言理论，也可进一步做有计算机辅助（CA）的比较研究。[12]

参考文献

1、徐通锵《基础语言学教程》2001年2月北京大学出版社，19-36页，178-237页[M]

2、北京大学计算语言学研究所《计算语言学文集》第4集“汉语语法研究所面临的挑战（98现代汉语语法学国际学术会议第一次全体大会宣读）陆俭明、郭锐”2000年12月，1-19页[C]

3、徐通锵《语言论--语义型语言的结构原理和研究方法》1997年10月东北师范大学出版社，295-442页[M]

4、邹晓辉“协同智能计算语言数据库的设计方法（2002年11月）--支持语言文字系统工程与全球语言定位系统的一个实施例”光明网论文发表交流中心[EB]

5、邹晓辉《第五届(国际)汉语词汇语义学研讨会论文集》“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 ”新加坡 2004年6月，281-286页[C] 光明网论文发表交流中心[EB]

6、张学文《组成论》中国科学技术大学出版社2003年12，44-56页，246-252页[M]

7、张志公《汉语辞章学论集》“汉语简论”(1996年人民教育出版社)，每一页[C]

8、南开大学《语言学论辑》“词语强制搭配的语义关系类别及其性质（作者：刘叔新）”（北京语言学院出版社1996年8月），1-17页[C]

9、喻云根《英汉对比语言学》北京工业大学出版社1994年12月，69-99页[M]

10、苏培成等《语文现代化论文集》“发挥汉语拼音在信息时代的作用（作者：冯志伟）”商务印书馆 2002 年 10 月，41-44 页[C]

尾注

[12] 例 1-5 的“（次广义的）字组”包括了“字、辞、块”或“词、词组、短语”等粗分形式。

[*1]文化基因进化发展的阶梯及层次的形式“类”。

[*2]语言信息标注、常识信息标注、（学科或领域）知识信息标注。

注1：邹晓辉“字的形式化定义--试论字本位理论的根基”[汉语“字本位”理论专题研讨会论文（短论之一）]2004年11月17日 光明网[EB]

注2：邹晓辉“字组的划分方法--试论字本位理论的功用”[汉语“字本位”理论专题研讨会论文（短论之二）]2004年11月27日 光明网[EB]

注3：邹晓辉“字与字组的关系--试论字本位理论的发展”[汉语“字本位”理论专题研讨会论文（短论之三）]2004年11月27日 光明网[EB]

汉语“字本位”理论专题研讨会论文

字与字组的关系

——试论字本位[1]理论的发展

摘要：

本文旨在：用字组显示字的具体意思。〝字的形式化定义〞、〝字组的划分方法〞和“字与字组的关系”三部曲[后者是前两者（〝姊妹篇〞）的进一步提炼或升华]用以表达：定义形式化、字组数字化和义项字组化[13]。根据〝音节总量控制模型〞（语汇数据库），每个字与字组，不仅在表中的id都是唯一的，而且，每个表的序号也是唯一的。汉语“字本位”理论与汉语（语汇）的形式化由此得以（间接）实现。

关键词：字本位、义项字典、字组用例

一、绪言

1、领域

本课题是“协同智能计算语言（理论模型）数据库（实现技术）设计方法[含：文本总量控制模型（GTCM）][2]”与“义项语汇典例（SVDE）的总量控制模型[GCM，涉及：GTCM 和 GSCM（音节总量控制模型[3]）]”的直接应用，即：以逻辑、数学和计算机工程的方法[4][5][6][7][8][9]尝试对汉语“字本位”理论做一些改进或优化，从而更好地指导“汉语（语汇）形式化”实践，属于：基础语言学与计算语言学（特别是计算语义学和语汇学部分）的交叉研究领域，进一步还会涉及：模式识别、自然语言理解和知识表达等具体研究领域[10][11][12][13][14]。

2、特殊性

在“字的形式化定义[15]、字组的划分方法[16]、字与字组的关系”构成的“三部曲”中，最后一篇是本研究的汇总报告，前两篇（姊妹篇）是基础，各篇分工是这样：“字的形式化定义”给出汉语形式消歧方法，即：定义形式化；“字组的划分方法”明确汉语形式划分标准，即：字组数字化；“字与字组的关系”给出汉语内容消歧方法，即：义项字组化。

3、重要性

“三部曲”有机结合一致论述“汉语（间接）形式化”新观点，旨在探索“巩固字本位理论的根基、明确字本位理论的功用、促进字本位理论的发展”的可能性，同时，验证“汉语形式化”的可行性，从而，在理论上，可为“汉语（间接）形式化”找到一条切实可行的新路；在实践上，可为设计优化“义项字典与字组用例”（义项大典和用例大全）提供一套新颖完善的方法和工具。

---

[13] 简称“三化”，即：定义表格化（采用双列表实现自然语言文字的间接形式化）、字组数字化和义项字组化。

4、研究途径

首先，在此借助“文本总量控制模型（GTCM）”序位的唯一性，从逻辑、数学和计算机科学的角度，论述定义形式化的唯一性，从而，奠定“汉语‘字本位’理论”体系形式化的基础；

接着，再借助“音节总量控制模型（GSCM）”序位的唯一性，从逻辑、数学和计算机科学的角度，论述字组数字化标准的唯一性，从而，奠定“汉语‘字本位’理论”应用工程化的基础；

然后，在前两项研究成果（即：形式化与数字化）的基础上，论述义项字组化的科学性。

5、局限性

本研究采取分步论证的策略，旨在分步考察“汉语形式化”的可行性，即：构成“三部曲”的三项研究，具体限制在“定义形式化、字组数字化、义项字组化”的论证规范以内。

6、基本假设

a、如果“文本总量控制模型（GTCM）”序位（如：第 4“进阶层式”）具有唯一性，那么，被称为字的汉语结构（形式）**类**，也就必然被唯一地限定在位于 GTCM 第 4“进阶层式”这一个特定论域或集合之中。至于具体的字的形式“**例**”则出现在 GTCM 第 4“进阶层式”的字表的 1...n 格的某个具体位置。

GTCM 第 4“进阶层式”这一个特定论域集合与 GSCM 第 1 表这一个特定论域集合一致。

b、如果“音节总量控制模型（GSCM）”序位[如：“2，3，...，m”（狭义）字组表的序号]具有唯一性，那么，被称为（狭义）字组的汉语结构（形式）**类**，也就必然被限定在位于 GSCM“2，3，...，m”表中。至于具体的字组的形式“**例**”则出现在 GSCM“2，3，...，m”系列字组表的“1...n”格的某个具体位置。

c、如果一个“字”有多个“义项”，那么，这个字与其多个义项之间则必然是“一对多”的关系。如果给每个义项设置一个唯一的序位标识号码（id）并与其在“音节总量控制模型（GSCM）” 2，3，...，m 一览表的（狭义）字组之间建立一一对应关系，那么，展示该义项的（狭义）字组，就可被视为“（与之一一对应的）**义项**本身”或“其显现”（两者由同义并列对应转换法则决定）。

与这个（狭义）字组一一对应的序位标识号码（id）就可以是该义项的“全权代表”，因为，遵循“同义并列对应转换法则”该义项、该字组、该序位标识号码（id）是等价的。其中，义项隐而不显（即：抽象的）；字组对识汉语的人传情达意；序位标识号码（id）对计算机传递信息。

d、如果“字本位、词本位、短语本位、...等各个（形式）本位”相互之间的**区别**主要表现在“结构形式”方面，那么其**联系**则主要表现在“概念内容”方面。也就是说，“字本位、词本位、短语本位、...等各个（形式）本位”都只不过是“内容本位”的某个则面或特例而已——表现为：某个具体的形式本位。

7、贡献

本研究发现并指出：汉语学界的（形式）“本位”之争实质上涉及形式本位与内容本位的关系。内容本位与形式本位通过“（字的）义项（解释）字组化”而实现融合。具体实现步骤如下：

a、字的形式化定义，通过定义形式化（形式消歧）而巩固汉语“字本位”理论的根基；

b、字组的划分方法，通过字组数字化（建立标准）而明确汉语“字本位”理论的功用；

c、字与字组的关系，通过义项字组化（内容消歧）而促进汉语“字本位”理论的发展。

上述三步骤的实现，不仅可为“汉语形式化”奠定（理论与实践两方面的）坚实基础，而且，也可为“自然语言间接形式化”提供可行的实施例。证明了“汉语形式化”具有可能性和可行性。

如果限定在形式范围，那么，不仅可从定性方面有效地排除“字与字组（形式化）定义”的概念歧义，而且，还可从定量方面唯一地确立“字组划分的标准”，这就论证了：汉语“字本位”观点（在汉语结构形式上）的合理性和重要性。如果从内容与形式统一的角度，那么还可进一步证明“字本位、词本位、短语本位、...等各个（形式）本位”，其实就是“（内容）本位”的某个则面或特例的形式化表现。总之，通过“字与字组的关系”的研究，把“本位”问题由理论概

念的抽象描述推进到了工程实践的具体操作的地步，进而，明确了对“形式本位与内容本位的关系”的认识——由“含混的合”到“清晰的分”再到“清晰的合”，从而开辟了义项字组化的汉语语汇形式化新路，为最终解决汉语形式化难题做了有益的探索。其优越性还在于为汉语形式化找到了一条标准化与个性化结合的新原理、新方法和新工具。

二、综述

汉语学界和中文信息处理领域的专家们，仍在被学界公认的“语义泥潭”所困扰。虽然以往和现行的汉语理论和中文信息处理原理及方法的研究途径仍然远离汉语形式化的康庄大道，但是，其中也不乏少数很有希望的具体研究，主要涉及：基础语言学和计算语言学以及人工智能等领域。

本研究比较关注：汉语“字本位”理论的“根基、功效、发展”与“汉语形式化”，尤其是字与字组的关系这一直接涉及“汉语语汇形式化”的课题。

在汉语“字本位”理论中，字与字组的关系，有两种表述形式：

1、“字”与“辞”或“块”之间的关系。

严格地说，无论是汉语“字本位”理论的“辞”与“块”这两种基本类型的（狭义的）字组，还是汉语“词本位”或“短语本位”理论的“词”、“词组”与“短语”，都存在一系列模糊区域或区分难题。试问：即使作为自然人的语言专家都难以区分的语言结构，如何让语言知识信息数据库都还很不健全的计算机系统来区分它们呢？看来，仅仅采用自然语言描述的汉语理论诸“本位”学说，都还难以而且可以说还都没有提出有效办法来很好地解决汉语或中文的“分词”难题。

2、“字”与“（二、三、四、…多）字组”之间的关系。

严格地说，这种划分（对中文信息处理而言）如果仅限于形式方面，那是相当科学而高效的。可是，一旦牵涉到内容或语义方面这个看似简单清楚的问题立即变得复杂起来。为什么会这样呢？

笔者认为：上述两个问题与学界公认的“语义泥潭”的问题有一个共同的本质——自然语言的多义性。可以说，这是造成自然语言理解的歧义性难题的根源。

因此，如果我们仅使用自然语言来论述这个课题，那么，字与字组（包括：核心字）的定义必然出现：内容的歧义性与形式的多样性，更不用说无歧义地论述“字与字组的关系”这样复杂的问题了。

首先，不同知识背景的人即使面对同一个对象或问题也都会有不同的认识（这是屡见不鲜的事情！），其次，自然语言本身即使对待同一个对象或问题也都会有不同的表述（这也是家常便饭！），最后，暂且不说“质疑者”提出的“挑战”一文是否找准了：汉语“字本位”理论的“七寸”或“短板”，仅仅从汉语“字本位”理论自身建设的角度考虑，也有必要采用“形式化”的方法，即：从“（现代）逻辑、数学和计算机科学”的方法和工具及实施例等一系列有效手段方面寻找：实实在在的有力支撑！[14]。

汉语理论的“词性划分”与中文信息处理的“自动分词”的一系列难点，一直是汉语语法与计算语言学的老大难问题。如果基于“字本位”观点，把形式上的“（狭义的）字组的划分方法”发展成为“（次广义的）字组的划分方法”，乃至“（广义的）字组的划分方法”，而且，最好能从“（现代）逻辑、数学和计算机科学”的方法和工具及实施例等一系列有效手段方面得到实实在在的有力支撑，那么，不仅可继续利用汉语“字本位”理论的知识积累与实践积累的现有成果以及相应的研究素材，而且，还可继续利用其他各种汉语理论的知识积累与实践积累的现有成果以及相应的研究素材（如，利用：基于“词本位”和“短语本位”的“词组的划分方法”的知识积累），更重要的是：可利用全面系统地间接形式化以后的汉语“字本位”理论在形式上可显著优于其它各种汉语理论与实践的高效率，首先，集中必要的人力、物力和财力，全面实现汉语语汇形式化（如，克服：基于“词本位”和“短语本位”的词组的划分方法的不足或缺陷，从而

[14]本研究正是在这方面做出的一个大胆而有益的尝试。到目前为止，已经初步完成了三项专题研究而且写出了专题论文（的初稿和修订稿）。现在，该是对整个研究进行归纳总结的时候了。

在解决上述“老大难问题”上面向前迈进一大步，让优化之后的汉语“字本位”理论的巨大作用，实实在在的体现出来！），进而，再合理布局分工合作，进一步全面推进“汉语（间接）形式化”的进程。

从“字与字组的关系”的角度来看，各种“本位”观点或学说，在“汉语形式体系”中，是如何定位的呢？以往的汉语理论与中文信息处理实践，在“老大难问题”面前，为什么显得那么乏力呢？其中的原因当然很多。但是，笔者认为：

1、局部的成功与全局的迷惑，是主要原因。

以下例 1-6 涉及“字、辞、块”或“词、词组、短语”粗分形式，可视为（次广义的）字组。

（一）局部的成功

例 1，汉语“字组”的两种组合方式[17]的探讨：

直接组合，以语序为主要手段把两个或两个以上的语言单位直接组合成一个较大的语言单位。如：组字成语、组语成句、组字成句。涉及“联合、修饰、陈述、动宾、补充”等组合关系。

关联组合，以起关联作用的虚字为主要手段把两个或两个以上的语言单位连接起来组成一个较大的语言单位。如：组字成语，涉及“并列、承接、递进、选择、转折、因果、假设、条件”等组合关系。

例 2，汉语“字组”的强制搭配[18]的研究：涉及对固定搭配和特定搭配（包括：控制性搭配与呼应性搭配）的语义关系类别及其性质的探讨。

例 3，英汉词汇对比[19]研究：涉及“广义词汇学（包括：语义学、词源学、词典学、修辞学）”与“狭义词汇学（词的性质、构成、意义及发展、词汇的构成及发展等规律）”的探讨。

例 4，在计算机上实行汉语的双轨制[20]的探讨：汉语“字组”及拼音形式与英语词语及音标形式之间，如何建立对应关系，是一个典型的人机协同研究或实验。

（二）全局的迷惑

例 5，语言理论探讨：对“（语音、语义、语汇、语法）四大要素”、“（语义、语法、语用）三个平面”乃至“（语音、语义、语法）三个方面、两层结构（表层结构与深层结构）”的研究。

最典型的就是面对语法与语义的相互缠绕的难题不知如何解决，从而导致了形形色色的各种各样的汉语语言观和方法论及其相应的具体的“只见树木、不见森林”的理论和方法。

例 6，语言结构研究：“字本位”、“词本位”、“短语本位”等形式本位，其具体语言结构形式，虽可各自称谓特定数量的概念或对象，但是，就“本位”学说的发展而论，均止步于：内容本位。

即使是汉语“字本位”理论，一旦深入到内容范畴，其问题或局限性也就立即暴露了出来。其它“本位说”（如：“词本位”、“短语本位”等形式本位）也不例外。

2、本研究的根据

本研究根据“文本总量控制模型（GTCM）”和 “音节总量控制模型（GSCM）”，以逻辑、数学和计算机科学的方法，确定“字与字组的关系”。

“字的形式化定义”已给出汉语形式消歧方法，即：定义形式化；“字组的划分方法[15]”也已明确汉语形式划分标准，即：字组数字化；“字与字组的关系”给出汉语内容消歧方法，即：义项字组化。

如果只考虑结构形式，那么，字与字组的关系是很清楚的。首先，只要给出字的形式化定义，进而，再明确字组的划分方法，包括给出字组的形式化定义和字组分类的统一标准，接下来也就自然能说清楚：

a、字本位，即：字，作为汉语的基本结构（形式）单位的合理性，事实上字在汉语形式化体系中具有不可动摇的根基地位；

[15]该文已从汉语（间接）形式化的角度提出把字组粗分水平（局部的成功）发展到字组细分的水平的建议！

b、基于字本位的汉语字组细分方法及标准，对汉语形式构造与解构的功用非常具体而直接。

c、在 a 和 b 所述的前两个专题研究基础之上，本专题将进一步尝试消除上述“全局的迷惑”。

随着本研究 a、b、c 三个专题完成并全面实施，不仅有利于汉语的计算机辅助（CA）研究、教学、应用乃至汉语间接形式化产品的生产和自动计算，而且，还将特别有利于电子词典（ED），百科全书（cyclopaedia，cyclopedia，encyclopaedia，encyclopedia）和海量常识知识库（CYC），数字图书馆（DL），专家系统（ES）等知识系统工程的改进或优化。

三、方法

为了探索“巩固‘字本位’理论的根基、明确‘字本位’理论的功用、促进‘字本位’理论的发展”的可能性，同时，探索“汉语（间接）形式化”的可行性。本研究提出了以下方法：

直接应用“自然语言（间接）形式化”的实用模型（GTCM 和 GSCM）探索“汉语形式化”的可能性和可行性，即：应用“逻辑、数学和计算机科学”方法，研究“字与字组的关系”，寻找汉语“字本位”理论与实践的优化方法，其特征在于：定义形式化、字组数字化、义项字组化。

其具体实施步骤如下：

在理论上，由“Ⅰ.字的形式化定义、Ⅱ.字组的划分方法（含：字组的形式化定义与字组划分的标准）、Ⅲ.字与字组的关系”三个专题组成；

如果说Ⅰ和Ⅱ两个专题着重研究汉语语汇的结构形式，那么，Ⅲ这个专题就将重点研究字的义项内容与（狭义的）字组的结构形式之间的关系。从而为“义项大典和用例大全”的设计优化，提供一套新颖完善的理论、方法和计算机工程模型。

在实践中，直接应用“自然语言（间接）形式化”的实用模型（GTCM 和 GSCM）作为具体探索“汉语（间接）形式化”的工具。

这是一条切实可行的路经，如：优化“义项字典与字组用例”（即：“义项大典和用例大全”的中试产品）的设计，同时优化“汉语的计算机辅助（CA）研究、教学、应用乃至‘汉语（间接）形式化’产品的生产和自动计算”的人机协作方案[16]。

1、Ⅰ和Ⅱ两个专题仅仅从形式方面确定字与字组的关系

（一）字的形式化定义（Ⅰ的回顾与提炼）

在**形式参照系**（GTCM 和 GSCM）中，规定具有**序位**唯一性（位于 GTCM 第 4“进阶层式”同时也位于 GSCM 第 1“进阶层式”）的（形式）**类**（方块形与单音节的**字**）为汉语的基本结构（形式）。见：图 1 和图 2。

图 1 是以“类”与“例”展示字的形式化定义的示意图。

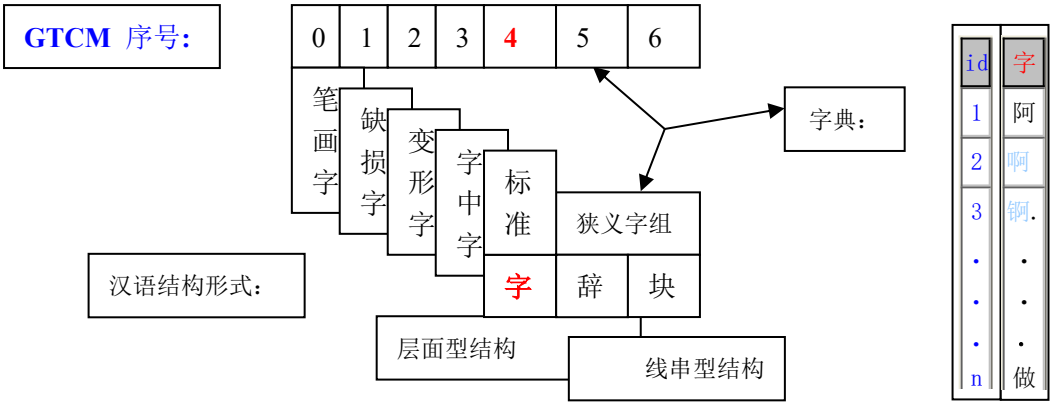


图 1

在图 1 中，笔者着重指出“字”和粗分“字组”的“**类**”以及“字”的“**例**”的直观**序位**。

形式化定义的“字”，特指：位于 GTCM 第 0-4“进阶层式”的“层面型结构”与位于 GTCM

[16] 这是生产式教学法及其系统工程方案课题的任务！

第 4-6“进阶层式”的“线串型结构”**迭交**于第 4“进阶层式”的**汉语结构形式**。

作为汉语“字本位”理论的“字”的“（形式）类”（定义）与“例”（字典中的字）应限定在“形式体系”的范围内（见：图 1 和图 2）。这样的“形式类”如“方块形”与“单音节”。至于，“字”的（内容）“多义项”特性，将通过“（字的）义项（解释）字组化”的方法而“形式化”[17]。

“形式体系”即“（纯）符号体系”，如：Unicode [统一的字符编码（标准）]和 ASCII（美国标准信息交换码）以及笔者拟制的 Z-ASCII（终极标准信息交换码[18]）[21][22]。

图 2 是在“形式化体系”之中进一步揭示字的形式类的示意图。

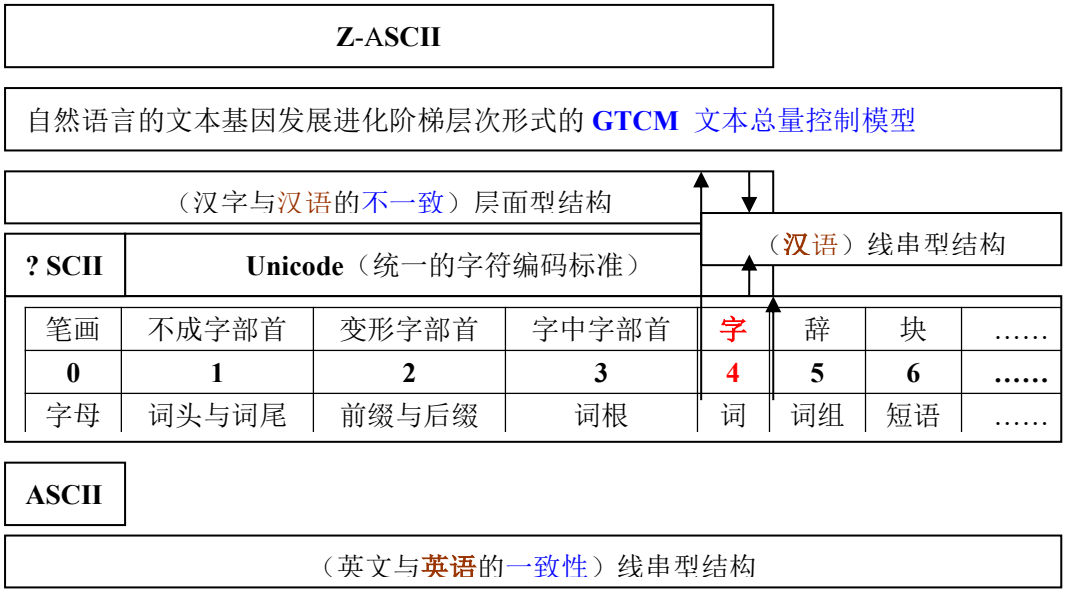


图 2

在图 2 中，笔者着重指出“字”和粗分“字组”在“形式体系”中的抽象序位。

GTCM 是采用计算机表达与处理自然语言的一种理想“形式体系”。Z-ASCII 的语言文字部分涉及 ASCII 与?SCII 或 Unicode 的汉字基本笔画以及标点符号、数字、外语字母、特殊符号等均位于 GTCM 第 0“进阶层式”，作为“基准参照系”，对 GTCM 第 1，2，…，m“进阶层式”的所有“组合结构”具有“测序定位”的作用[19]。

基于 GTCM，图 1 和图 2 给出了（字和字组）定义形式化的图解。

在图 1 和图 2 中，一方面，可从文字的角度，说明 GTCM 第 4“进阶层式”的“字”与第 0-3“进阶层式”的“基本笔画、（三级）偏旁部首”等“层面型结构”；另一方面，可从语言的角度，说明 GTCM 第 4“进阶层式”的“字”与第 5-6“进阶层式”的“辞、块”等“线串型结构”。

（二）字组的划分方法（II 的回顾与提炼）

图 3 是以“字组数字化”展示字与字组（在形式上）的关系的示意图。

---

[17] 这将在第（三）部分阐述，见：图 4、图 5 和图 6。

[18] 笔者拟制的 ChSCII（中文标准信息交换码）乃至理论上的 Z-?SCII（终极标准信息交换码），均可统一到 Z-ASCII 终极或中美（双文双语）标准信息交换码的体系之中。

[19] 这是“文化基因工程”课题的探讨范围。

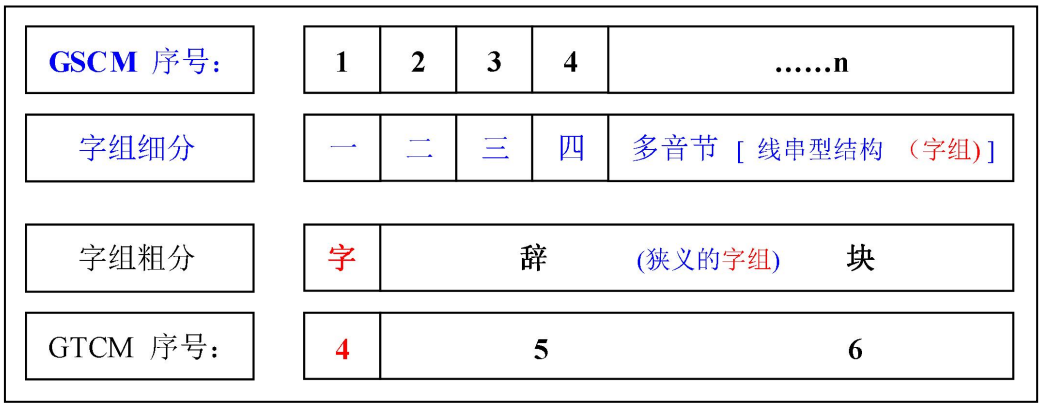


图 3

在图 3 中，笔者着重指出：（狭义）字组的划分方法，即：（狭义）字组数字化，包括：（狭义）字组的定义形式化和划分方法的标准化。十分明确地指出（狭义的）字组就是位于 GSCM 第 2，3，…，m“进阶层式”的“线串型结构”，其计量单位就是位于 GTCM 第 4“进阶层式”同时也位于 GSCM 第 1“进阶层式”的“字”。GSCM 第 2，3，…，m“进阶层式”的“2，3，…，m”数字指称的“字组”都包含对 GSCM 第 1“进阶层式”的“1”数字指称的“字”的重用。

在形式参照系（GTCM 和 GSCM）中，规定具有特定序位（位于 GTCM 第 5、6“进阶层式”或位于 GSCM 第 2、3、4、…m“进阶层式”）的一系列形式类（即：线串型字组或多音节字组）均为汉语的（狭义）字组（即：语汇层面除字以外的其它汉语结构形式）。

以下结合图 3 做详细说明：

在形式上限于语汇层面或 GSCM 范围的细分字组——次广义的，一方面，与 GTCM 第 4-6“进阶层式”（这一部分是粗分字组——次广义的）是同义并列的等价关系；另一方面，GSCM 细分字组又是 GTCM 粗分字组的数字化，GTCM 第 4-6“进阶层式”与 GSCM 第 1，2，…，m“进阶层式”之间，在语汇总量的可能性上也是一致的。图 3 就是这种一致关系的示意图。其中，GTCM 第 5、6“进阶层式”对应于汉语“字本位”理论所述的“辞、块”——粗分字组。GSCM 第 2，3，4，…，m“进阶层式”对应于汉语“字本位”理论所述的“二、三、四、…、多”字组——细分字组。

GSCM 第 2，3，4，…，m“进阶层式”体现了“字组数字化”或（形式化）字组划分方法的标准化，揭示了字与字组在形式上的关系，从而反映了汉语“字本位”理论的“形式化”功用，即：基于“字本位”的“字组细分”完全可借助计算机实施“数字化”自动分析或处理。

基于“（字与字组）定义形式化”的汉语结构描述的形式化处理和“字组数字化”的汉语语汇的形式化处理，原来仅仅采用自然语言（汉语）描述的“粗放型”汉语“字本位”理论与实践，就可以借助 GSCM 与 GTCM 进一步发展成为“精准型”汉语“字本位”模型，从而，为“（数理）逻辑、数学（计算）和计算机工程”方法的导入，奠定汉语“字本位”理论与实践在语汇层面上实际应用的基础[20]。

图 4 是以“逻辑、计算和工程”方法及实施例展示“义项字组化”的示意图。

[20] 图 4 用简洁的图形介绍了这种“实际应用”的概况。

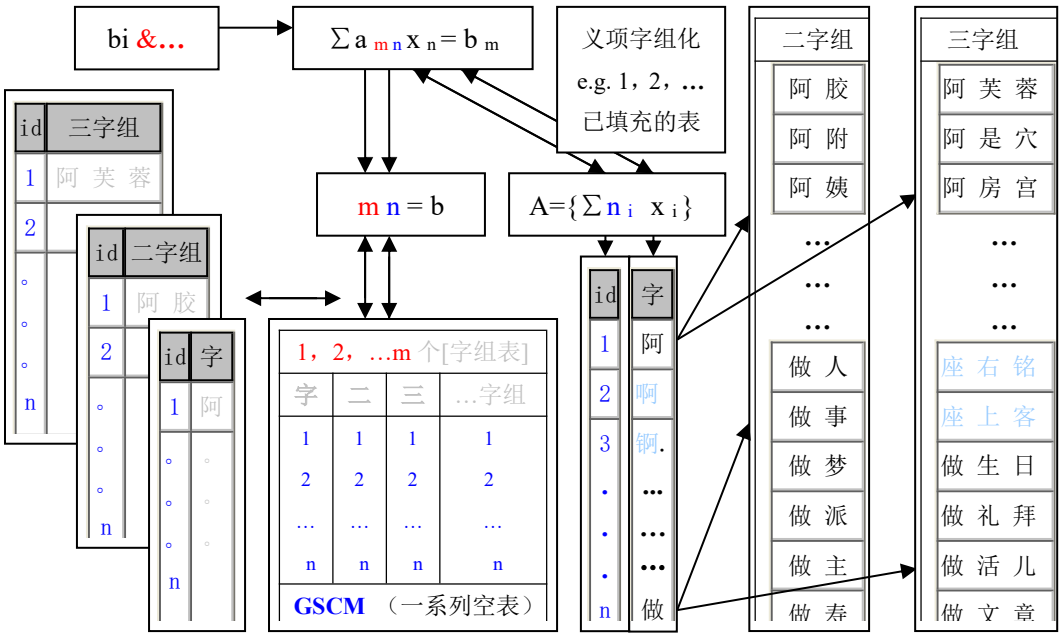


图 4

在图 4 中，笔者着重指出“（字的）义项（解释）字组化”不仅是一些想法和自然语言（汉语）的简单描述或说明，而且是已经得到“（数理）逻辑、数学（计算）和计算机工程”原理、方法和工具及实施例（乃至具体实践）有力支撑的事实。

以下结合图 4 做详细说明：

由图 4 可表示本研究提出把汉语“字本位”理论由“粗放型——仅仅采用自然语言（汉语）描述的原理”改进或优化为“精准型——同时采用（数理）逻辑（演绎与归纳以及枚举或穷举）、数学（计算）和（计算机）工程（实践）”方法描述的“汉语语汇形式化”理论模型和工程实例。

图左上方的“域（体）、环、群”三组代数公式和 GSCM 的“一系列空表”[21]展示了本研究的抽象部分；

图的右方的“字符多项式”[23]和“已填充的表”（计算机工程实施例）展示了本研究实例的直观部分，其中，例 1“阿”与例 2“做”两个字的“义项”解释，都是通过“二字组”、“三字组”、…的一系列具体的“（字组）用例”形式化的。图中未列举的例 3、例 4、…等其它实例与此同理。

2、Ⅲ这个专题同时从内容方面确定字与字组的关系

（三）字与字组的关系（Ⅲ专题的深入探讨涉及Ⅰ和Ⅱ专题的简单回顾）

在形式方面，字与字组的关系，在**形式参照系**的特定序位中，因明确定义了字和字组的形式类，故作为其汉语结构形式的具体含义与划分类型都非常清楚。

在内容方面，由于Ⅰ和Ⅱ专题在给出字与字组的形式化定义以及字组的形式划分方法的时候，回避了字的义项与字组的关系问题，因此，还有待Ⅲ专题对“字与字组的关系”做进一步的探讨。

笔者采用（字的）义项（解释）字组化的方法，进一步探讨：字与字组的关系。这实质上是：以“长字组（形式）”解释“短字组的义项（内容）”（转换或替代）方法的特例！。

该方法，一方面，在逻辑上遵循“外延扩大、内涵缩小”的法则；另一方面，在信息处理上遵循“同义并列、对应转换”的法则。

简单地说，就是：义项字组化，即：采用具体的字组作为解释字的具体义项的实际用例。这实际上就是：把“内容转述的过程”转换成为“形式替代的过程”（即：形式化过程）。

具体地说，则是：在形式参照系中，既要对具体的“字或字组”所在“进阶层式”一览表序号

[21] 这是“逻辑、数学、计算”原理、方法及工具的抽象分析课题的探讨范围。

（1，2，…，m），也要对具体的字的义项解释或字组转述或替代字组一览表序号（1，2，…，m），测序定位，从而确保同时实现形式与内容两方面的信息处理。见：图 4、图 5 和图 6。

图 5 是以“实例”展示字与字组（在内容与形式上）的关系的示意图。

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。

图 5

在图 5 中，笔者着重指出“具体的义项（内容）与具体的字组（形式）用例”是同义并列、相互替代的关系。

以下结合图 5 做详细说明

由图 5 可知，以一个字为“实例”可展示字与字组（在内容与形式上）的关系。

简单地说，该实施例就是对“义”这个字的“义项”（内容）解释直接采用含有“义”这个字的“字组”（形式）作为具体“（字组）用例”的。该例可视为以“义”这个字作为“核心字”在基于 GSCM“细分字组”数据库的“义项语汇典例（SVDE）”中分别向左（向心字组）或者向右（离心字组）自动查询“义”字的“义项”（内容）或“（字组）用例”（形式）的“搭配限制信息”。

“搭配限制信息”，除了上述直接制约的途径之外，还有间接制约的途径，具体涉及三种形式，即：a、语言文字信息标注（形式部分，涉及字组的“生成、采集、比对、转换”技术信息；内容部分，涉及语言理论与实践的各个方面[22]）；b、通用常识信息标注；c、专家知识信息标注[23]。

基于 GTCM 与 GSCM 的（特定）协同智能计算系统，可成为关于“义”这个字的已知各种“搭配限制信息”的“集大成者”或“一字之师”——相当于特殊的“电子字典和词典（ED）”、“常识系统（CYC）”和“专家系统（ES）”或相应的计算机辅助（CA）“汉语教学应用系统”。一旦“该（特定）系统”遍历了所有常用字的（形式）搭配限制信息，即可实现汉语语汇形式化，从而可为中文信息处理的自动化乃至智能化奠定重用该（特定）系统已掌握的搭配限制信息基础。

图 6 是以“实例”展示字与字组（在内容上）的关系的示意图。

概念本位： 箭头表示：**义项**呈**收敛**趋势的方向

本真信息的总称：	序位本义	本义	义	（同一名称）称谓信息的本质
形式信息的总称：	广义文本	文本	文	（同一名称）指称：信息现象

22 见“语义分析新方法”与“内容消歧新方法”。

23 见“电子字典和词典（ED）”、“常识系统（CYC）”和“专家系统（ES）”分析。

图 6

在图 6 中，笔者着重指出义项（内容）收敛与用例字组（形式）长度增加两种变化呈反比。

以下结合图 6 对做详细说明：

由图 6 可知用一组词语或字组表达同一概念的实例展示字与字组（在内容上）的关系。

简单地说，用“文、文本、广义文本、…”这一组词语或字组表达同一概念，即“信息现象”。也就是说，“文、文本、广义文本、…”这一组词语或字组都是用于称谓或指称“形式信息”的。在此，“文”这个字的“义项”随着“（字组）用例”的形式变化（即：字组延长或字数增加）而发生相应变化（即：义项收敛）。用“义、本义、序位本义、…”这一组词语或字组表达同一概念，即“信息的本质”。也就是说，“义、本义、序位本义、…”这一组词语或字组都用于称谓或指称：“本真信息”的。在此，“义”这个字的“义项收敛”与前面同理。 其它用例的道理也相同。

综上所述，图 4、图 5、图 6 强调（字的）义项（解释）字组化的“内容与形式”转化原理。图 1、图 2、图 3、图 4、图 5、图 6 在内容与形式上，全面阐述字与字组的关系。

字与字组的关系，在形式上，就是位于 GSCM 第 1“进阶层式”的“字”的“（形式）类和（枚举）例”与位于 GSCM 第 2，3，…，m“进阶层式”的“字组”的“（形式）类和（枚举）例”之间的关系，进一步涉及“字”的“义项”（内容）与具体的“字组”（形式）之间（在内容与形式上）的关系，以及（仅仅作为形式的）字与字组（在内容上）的关系。

四、结果、结论和议论

1、结果

GTCM 和 GSCM 在本研究中的应用，实质上是“自然语言形式化”（方法）在“汉语（语汇）形式化”理论与实践上的具体尝试，本研究“三部曲”试图证明：不仅从理论上可以理清汉语“字本位”理论在**字与字组的关系**上的一系列仅采用自然语言说不清楚的问题（必然存在的歧义），而且，还可得到“汉语形式化”（包括：理论自身描述的形式化与在改进或优化的汉语“字本位”理论指导下的汉语语汇形式化工程实践）的“三级跳”方法和工具及其实施例的可行性方案，即：

通过给出“字的形式化定义”而坚固：汉语“字本位”理论的根基；

通过提供“字组的划分方法”而突出：汉语“字本位”理论的功用；

通过阐明“字与字组的关系”而加速：汉语“字本位”理论的发展。

2、结论

上述结果中，（字的）义项（解释）字组化，是：本研究在试图进一步明确“字与字组的关系”的过程中的一个（既带有几分幸运又带有几分必然）的重大发现。

这个发现是实质性的。它使“汉语语汇形式化”可能性与可行性得到了完全证实，因此，也必将显著改变或大大加速“汉语形式化”的进程。

事实证明：基于 GTCM 和 GSCM 的“汉语（语汇）间接形式化”，不仅具有理论上的必要性和可能性，而且，还具有实践上的必要性和可行性，特别是：汉语语汇的静态（形式化）处理与动态（形式化）处理，已经被实践和事实证明是切实可行的，具体成果体现在“义项字典与字组用例”之中[24]。

本研究不仅为“汉语理论的概念体系的形式化”和“汉语（语汇）形式化”做出了十分重要而有益的尝试，而且，还验证了基于 GTCM 和 GSCM 的“汉语（语汇）形式化”系统工程方案，是切实可行的。

为今后的推广普及奠定坚实的理论基础和两大系统工程（即：基于 GTCM 和 GSCM 的汉语信息处理与汉语教学体系）基础，如给予及时的扶持便可让本研究提出的具体项目能早日完善。

本研究指出：“巩固‘字本位’理论的根基”、“明确‘字本位’理论的功用”和“促进‘字

[24] 见：“汉语义项字典”与“双语字组用例”以及“多语通用字组用例”几项自然语言处理的工程化课题！

本位’理论的发展”，不仅具有可能性，而且，具有可行性。

也就是说：基于 GTCM 和 GSCM 的“汉语形式化”方法，不仅可在汉语理论的自身建设方面做到：（使所有的汉语结构的）“定义形式化”，而且，还可通过“字组数字化”和“义项字组化”，全面推进“汉语形式化”的步伐。

本研究证明，首先，由“字本位”确立的“字与字组”的“结构形式”是容易识别的；同时由“字本位”确立“义项字典与字组用例”的“知识内容”是可有针对性重用（理解或表达）的。

义项表达的形式化（即：字组化）和标准化（即：格式化、代码化、数字化），作为本研究的特点是独特而高效的。

基于这种“形式化”和“标准化”的思路和方法在文本总量控制模型（GTCM）和 音节总量控制模型（GSCM）中从应用的角度看字与字组的关系是“义项大典”与“用例大全”的关系[25]。

一方面，从语言形式上看，“字”是构成一切“字组”的“基本结构单位”；

另方面，从语言内容上看，“字”又必须借助“字组”的“形式展示义项”。

本研究不仅在（基于汉语“字本位”理论“字组的划分方法”的）“字组粗分”与“字组细分”之间建立了“规范的形式化体系”或“相互转换的理论模型”，而且，还为这种理论模型提供了：“逻辑、数学和（计算机）工程（实践）”的基础。

也就是说，仅采用自然语言（汉语）描述的“粗放型”的汉语“字本位”理论，经过适当的改进或优化，将可能成为：得到“逻辑、数学和计算机工程实践”有力支撑的“精准型”的汉语“字本位”理论——准确地说是以“字本位”观点为基础的“汉语形式化”理论——包括理论的形式化与汉语的形式化。

3、议论

下面从汉语理论和汉语实践（举例说明）两个方面讨论有待进一步系统思考的问题。

a、理论的形式化

由于字与字组的关系在汉语“字本位”理论体系的总体框架中具有非常重要的地位，因此，有必要分层次地直观展示（改进或优化的）汉语字本位理论体系中诸学科之间的基本关系。

图 7 是（改进或优化的）汉语字本位理论体系中诸学科之间的基本关系的示意图。

语言	语用、文体或章法	话题-说明（句群、段落、篇章）
	语法、翻译	主语-谓语（句子）
	语汇、修辞	字组（内容与形式）或词组
	语义（义项分析）	语言基本结构单位（内容方面） （迭交的）字（汉语）或词（英语）
	语音（音素分析）	语言基本结构单位（形式方面） 单音节-汉语的（音）字 单、双、多音节-英语的词（音）
文字	文字（部件分析）	（汉）（形）字（部件：笔画、偏旁部首） （英）词（形）（部件：字母、语素）
	符号（Z-ASCII）	（汉）笔画 （英）字母

图 7

由图 7 中“语音（e.g 汉字的单音节或英词的多音节）、语义（语言基本结构单位 e.g 汉语的字或英语的词）、语汇（e.g.字组或词组）、语法（e.g.主语-谓语）、语用（e.g.话题-说明）”的关系，可看出“字”的根基地位和构成“字组”的基本功能以及（改进或优化的）汉语字本位理

[25] 字与字组或词的关系是“义项大典”与“用例大全”的关系。

论体系。

b、语言的形式化

由于字与字组的关系在汉语（形式化）体系的中重要地位，因此，有必要分层次地直观展示基于（改进或优化的）汉语字本位理论体系而建立的“汉语（语汇）形式化”方法的这三个基本步骤，即：字与字组的定义（间接）形式化、分类（间接）数字化和义项（直接）字组化。

图 8 是“汉语（语汇）形式化”（基本步骤）示意图。

	汉语（形式化）	英语
语汇	单音节-字；双、多音节-**字组**（数字化）	混音节-词；双、多音节-词组
语义 义项分析	（迭交的“**字**”的各个）**义项**（字组化） **字** [语言基本结构单位（内容）]	 词 [语言基本结构单位（内容）]
语音 音节分析	单音节-**字**；双、多音节-**字组**（形式化） 语言基本结构单位（形式）	单、双、多音节-词 语言基本结构单位（形式）

由图 8 中“字（单音节）与字组（双、多音节）”的关系，可看出“字在汉语（语汇）形式化”进程中的作用。字与字组的定义（间接）形式化、分类（间接）数字化和义项（直接）字组化，在汉语（语汇）形式化过程中，实际是因为迭交的“**字**”而“迭交”在一起的。比较而言，英语的“词（混音节——单、双、多音节）与字组（双、多音节）”的关系，既简单又复杂。

c、形式本位与内容本位的关系

由于理论的形式化与语言的形式化，不仅可从汉语“字本位”理论与实践两方面优化；而且，还可从（普通语言学的）语言“本位”理论与（自然语言理解的）语汇形式化实践两方面优化，因此，笔者进一步提出“形式本位与内容本位的关系”的深层次问题。具体讨论见以下实例。

仅就汉语语汇而论，字与（狭义）字组在“结构形式”与“概念内容”之间的相互关系如下：

就合成概念及其内涵基本概念（**内容**）而论，字与字组（**形式**）之间可有如下基本关系：

（1）可分的字组，即：松散合成的字组，如要深究其中蕴涵的概念，则**合不如分**。

可分字组：意义、事物、.....

合不如分：意与义、事与物、.....

否则，有的概念可能被忽略。如：“意义（= meaning）”，则“意”与“义”两个概念就无法直接翻译（人都难，何况计算机！）。

（2）不可分字组，即：紧密构造的字组，无须深究其中蕴涵的概念，故**分不如合**。

如：序位本义、广义文本、.....

否则，概念难以表达清楚。如：融智学有一对基本概念，虽然采用“字、二字组、四字组、…（语言的结构形式不同）”来表达均可，但是长度适中的那一对字组，表达更清楚（更易排除歧义！）。

例 1：这一对**基本概念**（**内容**）及相应的那一系列**字组**（**形式**），分别（静态）表述如下：

义、本义、序位本义（范畴相同，义项收敛），表示融智范畴体系的本质，属：本真信息。

文、文本、广义文本（范畴相同，义项收敛），表示融智范畴体系的现象，属：形式信息。

例 2：这一对**基本概念**（**内容**）及相应的那一系列**字组**（**形式**），组合（动态）表述如下：

a、**文**以载**道**。（在此，设：“义=道”，即：两字表示同一范畴）

b、“文本”表示“本义”。

c、“广义文本”用以表示“序位本义”。

在形式上，a 句，适合“字本位”，如：“文”和“义=道”两字既可以表示两个与 b 句相同的一对基本范畴，又可以表示两个与 c 句不完全相同的一对范畴。b 句，好像适合“词本位”，如：“文本”和“本义”在此既可以表示与 a 句相同的一对基本范畴，又可以表示两个与 c 句不完全相同的一对范畴。c 句，非常适合“短语本位”，如：“广义文本”和“序位本义”在此明

确表示一对**大的组合概念**——蕴含“物、意、文”的“广义文本”和仅蕴含“义”的“序位本义”。

在内容上，a、b、c 三句，都适合“内容本位”，如“文、文本、广义文本”三者虽然都可以用来表示“形式信息”这个很大的范畴（属：同一个很大的概念）但是表达不精细（蕴含歧义）。由此可见，“字本位”、“词本位”、“短语本位”等“**形式本位**”的必要性和局限性[26]。

如：a、b、c 三句，“文、文本、广义文本”，在各自的句子中，分别都是基本结构（形式）。这样看来，“字本位”、“词本位”、“短语本位”等“**形式本位**”似乎都成立，各有各的适应条件。

例 3：以下两个句子，表示**相同的**意思或用意：

a、融智学是一门讲述“**文**以载**道**”的学问。

b、融智学是一门讲述“**文本**表示**本义**”的学问。

c、融智学是一门讲述“**广义文本**表示**序位本义**”的学问。

d、融智学是一门讲述“**广义形式信息**表示**序位本真信息**”的学问。

……

例 4：以下两个句子，表示**相同的**意思或用意：

a、**文**与**义**是**融智**这一核心**活动**的两方面。

b、**文本**与**本义**是**融智**这一核心**活动**的两个方面。

c、**广义文本**与**序位本义**是**融智**这一核心**活动**的两个方面。

d、**广义形式信息**与**序位本真信息**是**融智**这一核心**活动**的两个方面。

……

试问：能简单根据“字、二字组、四字组、…”区分出上述句子中哪一对是基本结构单位吗？

（3）两可的字组，即：半松半紧的字组，深究其中蕴涵的概念，均可，即：**可分可合**。

如：形与式、形式，序与位、序位，.....

由“（1）**合不如分、**（2）**分不如合、**（3）**可分可合**”三种情形可见，“字组”结构的“形式”与“内容”一旦交织在一起，汉语结构（形式）本位（如：“字本位”、“词本位”、“短语本位”…）诸说就恰似“盲人摸象”各执一端的实际情形。

为简明扼要地指出汉语学界关于语言基本结构的“本位”之争存在的问题，本文特将现有的“字本位”、“词本位”、“短语本位”和“小句本位”乃至“复本位”等统统归入“形式本位”的范畴。与之对应，提出“内容本位”（如：概念本位，关系本位）的观点或学说。

这样，就可以说，当前汉语学界关于语言基本结构的所谓“本位”之争，实质上涉及学界对“形式本位与内容本位的关系”的认知程度！。

“形式本位”，特指：以某种符号形式（如：汉字或英词）作为“基本结构单位”；

“内容本位”，特指：以某种思想内容（如：范畴或概念）作为“基本思想元素”。

这实质上就是以逻辑分析为指导的“语言分工”说，如：“汉字或英词”表达“范畴或概念”；“字组或词组”以及“句群或句子”表达“（范畴或概念之间的）关系”。

笔者主张“形式本位”与“内容本位”结合的观点。因为，只有这样，才能看清汉语“这头大象”的全貌。

众所周知，凡是实际问题，其内容与形式都是结合在一起的。理想的形式化或标准化都是在特定的条件下构造或形成的抽象或直观的模型（无论它多么实用）。可以说，这些条件通常也都是为了便于分析或思考问题而创设的。因为，我们的感官能力、记忆能力和思维能力等都非常有限，一旦超越其极限（主体自身发展的知识和技能局限）就必然造成识别、理解与表达上的含混不清。

所以，仅仅依靠自然语言还不够，必须借助相应的方法和工具，如：现代逻辑学与数学乃至计算机科学（理想的形式化或标准化手段就随之得到了发展和推广），才能把复杂的问题表述清

---

[26] 各个“**形式本位**”仅仅适合满足特定形式的“句类、句型、句式、句例（具体的句子）。

楚。

综上所述，如果仅在形式的范围考虑，那么，不仅在定性方面可排除字与字组概念的歧义，而且在定量方面也可确立字组划分的标准，进而在结构形式上论证“字本位”的合理性和重要性。

几乎所有的复杂问题，都是内容与形式的统一。“字本位、词本位、短语本位、...”等“形式本位”其实都只不过是“内容本位”的某个则面或特例。如果要真正发展汉语的“字本位”理论，就不仅要应用“形式化”的方法和工具，而且还要发现或理解“形式化”方法和工具背后的机理。这就是本文讲述的一个重要观点或核心思想，即：在参照系（即：GTCM 和 GSCM）中，实现：（理论上，字与字组）定义形式化，（实践上）字组数字化和（字的）义项（在形式上）字组化。

参考文献

1、徐通锵《语言论--语义型语言的结构原理和研究方法》1997年10月东北师范大学出版社，295-442页[M] 徐通锵《基础语言学教程》2001年2月北京大学出版社，19-36页，178-237页[M]

2、邹晓辉“协同智能计算语言数据库的设计方法（2002 年 11 月）”[J]《潜科学》第 32 期

[EB]

3、邹晓辉 “义项语汇典例（SVDE）的总量控制模型（2004 年 6 月） --人机协作对采用汉语注释的语义词汇典例进行计量分析”284 页[C] 《第五届(国际)汉语词汇语义学研讨会论文集》

[J]《潜科学》第 32 期

光明网论文发表交流中心转载[EB]

4、朱志凯《逻辑与方法》1995 年 8 月第一版，人民出版社 3-32，225-287，229-304 页[M]

5、北京大学数学力学系几何与代数教研室代数小组《高等代数》1978 年，人民教育出版社 1-49，102-149，376-398 页[M]

6、熊全淹《近世代数》1978 年 8 月第二版，上海科学技术出版社 15-120 页[M]

7、中国人民大学数学教研室《线性代数》1983 年第一版，85-138 页[M]

8、[美]David M. Kroenke 《DATABASE PROCESSING——Fundamental，Design & Implementation（Seventh Edition）》施伯乐等译《数据库处理——基础、设计与实现》2001 年 3 月第一版，电子工业出版社 170-246，334-489 页[M]

9、康博创作室《SQL Server 2000 数据仓库设计和使用指南》2001 年 4 月第一版，清华大学出版社 14-36，49-69，113-230 页[M]

10、陈肇雄主编《机器翻译研究进展》1992 年 8 月第一版，电子工业出版社 1-564 页[C]

11、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》1998 年 11 月第一版，清华大学出版社 1-516 页[M]

12、北京大学计算语言学研究所《计算语言学文集》第 4 集，2000 年，1-254 页[C]

[EB]

13、黄河燕主编《机器翻译研究进展》2002 年 11 月第一版，电子工业出版社 1-282 页[C]

14、苏培成等《语文现代化论文集》2002 年 10 月，商务印书馆 1-364 页[C]

15、邹晓辉“字的形式化定义--试论字本位理论的根基”[汉语“字本位”理论专题研讨会论文（短论之一）] 光明网[EB] 2004年11月17日 [J]《潜科学》第38期

16、邹晓辉“字组的划分方法--试论字本位理论的功用”[汉语“字本位”理论专题研讨会论文（短论之二）] 光明网[EB] 2004年11月27日 [J]《潜科学》第38期

17、张志公《汉语辞章学论集》“汉语简论”(1996 年人民教育出版社) [C]

18、南开大学《语言学论辑》“词语强制搭配的语义关系类别及其性质（作者：刘叔新）”（北京语言学院出版社 1996 年 8 月），1-17 页[C]
19、喻云根《英汉对比语言学》北京工业大学出版社 1994 年 12 月，69-99 页[M]
20、冯志伟“发挥汉语拼音在信息时代的作用”商务印书馆 2002 年 10 月，41-44 页[C]
21、邹晓辉“论影响人类未来的五大系统工程之间的关系”《熵.信息.复杂性（Entropy Information Complexity）》[J]第 86 期 [EB]
22、邹晓辉“一种知识信息数据处理方法及产品”2000 年，G06F163[C]知识产权出版社
光明网论文发表交流中心[EB]
[EB]附图 1-6
23、张学文《组成论》中国科学技术大学出版社 2003 年 12 月，44-56 页，246-252 页[M]“字符多项式与表格数学”[J]《潜科学》第 39 期（转载） [EB]

字本位与汉语形式化[27]

在“字、辞、块、读、句”与“字、二字组、三字组、…、多字组”的结构划分过程中，汉语“字本位”理论，强调：“字”是汉语的“基本结构单位”。其“核心字、两点论、语义句法”给笔者印象极深。出于“探寻汉语思维特点”的“好奇心”和“寻找改进汉语理论和中文信息处理以及计算机辅助汉语教学的新方法”的“强烈愿望”，笔者采用“字本位”的上述两种“结构划分”细化了“一种知识信息数据处理方法及产品（珠海邹晓辉的发明 2000）”的汉语部分。

本文主要论证“字本位与中文信息处理”方面的探索成果与研究心得，即：在完善“两表”的基础上，用“两表”为“参照系”进一步解析“字与字组的关系”[28]。

8.3.1 “字本位”与“两表”

基于“字本位”而构造的“汉语语汇数据库”，用事实证明了汉语“字本位”理论的优越性。

基于“字本位”而确立的“字与字组的关系”，在“两表”中可得到“形式化”体现。

由文本总量控制模型（GTCM）“4，5，6”分表构成“汉语（的字和基于字的）字组粗分模型”（见：图 1）；由音节总量控制模型（GSCM）“1，2，3，…，m”分表构成“汉语（的字和基于字的）字组细分模型”（见：图 2）。

以下的探讨所述的“两表”特指“汉语字组粗分模型”与“汉语字组细分模型”。

下面用“两表”作为解析“字与字组的关系”立体坐标从“字内信息、字间信息、字外信息”三个方面，探索“字本位”与“汉语形式化”结合的新路。

“形式化”通常是就“形式语言”、“程序语言”或“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。就此而论，“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“中国标准信息交换码”（ChSCII）。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“0-4 分表”处理。如果这个工作得到国家支持，我们就可早日开发出基于 ChSCII 的计算机中文输出输入系统（ChBIOS）和中文字库（ChFONTS）。由于 ChBIOS 与现有的英语 BIOS 兼容且平级因而可用汉语直接控制，ChFONTS 与现有的汉语 FONTS 兼容且与拼音字库平级因而也可用汉语直接控制。

在此基础上“字间信息”由 GTCM“4-6 分表”处理。如果这个工作得到普及，人们就可早日开发出基于 ChSCII 和 ChFONTS 的能“直接在计算机底层用汉语思考与表达的软件开发平台”。如

---

27 注：本文的修订稿被《字本位理论与应用研究》一书收录为 8.3 的压轴篇章，见 8.“字本位与中文信息处理”。
28 其中，“三化”的行文改进或优化，见《字本位与中文信息处理的基础》（即：邹晓辉著“融智学导论”）。

能完成上述两步，那么我们才可以说“中文信息处理”真正上了一个大台阶。由于语言处理必然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“5-12 分表”处理的过程。也就是说，如果能完成上述三步，那么，我们才可以说“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以英语公开，所以，除了解决汉语本身“字与字的语法接续问题”之外，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”——由“GTCM 的 5-6 分表”处理的“释义字组”，也就成了本文关心的一个重要方面。搞清楚“字与字组的关系”有利于解决上述这些实际问题。就语汇而论，GSCM“1-m 分表”与 GTCM“4-6 分表”总量相等。“用汉语思考与表达的中国人”与“用英语思考与表达的外国人”能否有平等地位？关键在于对表达“对象、概念、关系”的“释义字组”能否掌握到位？

就“字与字组的关系”而论如果笔者从语言事实中发现的“迭交原理”、“等价原理”、“基本组字公式”和“基本字组方阵”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，（改进或优化之后的）汉语“字本位”理论的优越性必将举世公认。那时，基于汉语且兼容英语的高性能计算机和中文操作系统（ChOS）也才有可能出现。ChOS 与英文操作系统兼容且平级从而可用汉语直接控制，区别于基于英文操作系统的“汉化”中文操作系统。

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）			

8.3.1.1“两表”

[1] 字组粗分模型

图 1 是 GTCM 示意图。

(1) 简述 13 个分表

在图 1 中，一览总表展示了汉语的 13 个分表的形式“类”。各个分表均以“数字与文字”为计算机前台接口，以“整型与字符串”为计算机后台“数据结构”。由此建立“汉语表格化”计算语言形式体系。本文重点探讨汉语语汇“字、辞、块”3 个分表，其它 10 个分表仅作简要介绍。

在图 1 中，对汉语而言，0，1，2，3，4 五个分表记录的“字内信息”可计算，其“符号”总量有限[29]；4，5，6 三个分表记录的“字间信息”可计算，其“语汇”总量也相对有限，广义的词汇学还包括字典词典学、语义学、修辞学。7，8 ，9，10，11，12 六个分表记录的“字外信息”可计算；其总量具体到目标用户日常处理能力范围也是相对有限的。

这个自然段的术语，仅供参与“协同智能计算系统”设计和数学分析的读者提供，其他读者可不读此自然段。0 分表的元素“集合”被命名为“子全域”，其“元素”数目虽极为有限却是构成后续 1，2，3，4，5，6，7，8，9，10，11，12 分表的所有“元组”——被命名为“超子域”。“子全域”的特性类似“生物基因”(ATGC)，故被视为“文本基因”。0，1，2，3，4，5，6，7，8，9，10，11，12 被命名为文本基因及其组合的“进阶层式”，它们是“子全域”与“超子域”相互连接的纽带或具体场所。“超子域”的复杂程度，在微观上取决于“基因文本”的分布状态——根据“迭交”原理可解析，其计算、统计、分析，以“子全域”为基准参照系；在宏观上则取决于“进阶层式”的具体层级，其计算、统计、分析，以具体的“进阶层式”为应对参照系。

[29] 8.3.5.1 解析“字内信息”将做进一步分析（见实例 1 和图 9）。

“超子域”的“可解析程度”视其组配结构涉及的“子全域”及其“元素”交叉重叠程度而定。

(2) “字、辞、块”

由于本文探讨汉语语汇，所以，4，5，6，即：“字、辞、块”三个分表是图 1 的焦点。图 2 将对“辞、块”做进一步细分，以便计算机自动化处理。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。

在图 1 表中笔者提出“音字”的概念和“混音节”的称谓，同时，把“形字、音字、实字、虚字”并列，突出“音字”与“字音”，“混音节”与“多音节”，“形字”与“字形”的区别，旨在说明各个术语强调的重心、关注的焦点以及主要研究对象均不同。采用部分新词或新概念，一方面，是为使表达更到位；另一方面，也是给后面将要介绍的“迭交原理”、“等价原理”和“基本组字公式”以及“基本字组方阵”做准备。如，例 1 的“用字”+“解字”=“释辞”，例 2 的“前字”+“后字”=“二字组”。例 1 与例 2，虽可表达同样公式，但精确程度却有所不同。“形字、音字、用字、解字、释辞”等术语目前虽是笔者的见解，但因能突出汉语与英语在结构形式或表现形态上的区别，故采用[30]。

(3) 单音节与混音节的区别和联系

同样位于 GTCM 第 4 分表（见：图 1）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（形）字是基于笔画的“层面型结构”[31]；（英文）词（形）是基于字母的“线串型结构”[32]。另方面，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。

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 2 是 GSCM 示意图。

(4) 字、辞、块与词、词组、短语

在 GTCM 中位于同样进阶层式（见：图 1）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”形式或模型是难以发现的。

[2] 字组细分模型

(1) 字组“细分”形式或模型概述

在图 2 中，一览总表展示了汉语语汇和英语词汇 1，2，3，…，m 个分表的形式“类”。这里，仅限于形式比较。“细分”就是以“字”作为汉语“线串型结构”的“起点”和其它诸“节点”，即：作为“度量”（各级）“字组”的“基本尺度”（基本结构形式单位）。“两表”结合使用，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 中就表现出了实质性差异。其区别的标志就在于是否包含“单音节”，图 2 使其“一目了然”。

(2) 仅就汉语语汇而言，GSCM 与 GTCM 的总量相等

GSCM 的细分与 GTCM 的粗分只是汉语语汇“子类划分”的改变，虽然各“子类”的“例”在归属上会有所不同，如：GSCM 有“1，2，…，m”多个细分“表”而 GTCM 只有“4，5，6”三个粗分“表”，但是，就汉语语汇而言，GSCM 与 GTCM 在总量上却都是相等的。

---

30 随着论述的深入，将有相应的说明。对此不习惯的读者可暂时不管它——采用旧词或旧概念虽不精确但也有近似效果。欢迎读者或学界同行提出宝贵意见！

31 2.字的定义形式化与 8.3.5.1 中解析“字内信息”将做进一步介绍和分析。

32 2.字的定义形式化与 8.3.5.1 中解析“字间信息”也将做进一步介绍和分析。

(3) 图 2 的焦点——“（汉）字”与“（英）词”比较

在图 2 中，汉语的（音）字与英语的（单）词”在 GSCM 中地位不同，因此这两种语言的基本结构（形式）单位的区别是显而易见的（形态上也不同）。第一，静态区分：字仅仅限于 GSCM 第 1 一分表，理论上讲词可位于 GSCM 第 1，2，…，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”，两者均属“线串型结构”。

(4) 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、短语之间的微妙关系，在 GSCM 和图 2 中得以展现[33]。具体要点如下：

首先，如果就语言结构（形式）而论，那么，“字”与“词”之间的区别是最根本的。因为，同样位于 GSCM 第 2，…，m 多个分表的（汉语）字组（辞或块）与（英语）词组或短语，其相互之间的区别皆由其基本构成单位的“字”与“词”的区别而派生，各自的组配法则也有区别。

其次，英语“词”（混音节，含：单音节、双音节、多音节）与英语“词组或短语”（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可被后者包含，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”之间却有空格（这被计算语言学界认为是英语的一个优点）增加了“英文信息处理”的识别标识。

最后，如果仅限于 GSCM 来看，那么，很显然，构成汉语“字与字组”的“音字”之间无空格（这一点被计算语言学界认为是汉语的一个弱点）增加了“中文信息处理”的困难。但如果考虑 GTCM 可能提供的字内信息，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）。

(5) 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 2 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“字、辞、块”的关系和“（英）词组或短语”与“辞、块”的关系。

(6) 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）奠定基础，即：表格化、数字化[34]。

8.3.1.2 字与字组的关系——兼谈“字”与“词”的区分

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

[1] “三化”（由 8.3.2，8.3.3，8.3.4 三节分别逐一介绍）

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处理汉语的一条快捷方式。

[2] “三注”（ 由 8.3.4 节后半部分介绍）

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”）。这是计算机辅助处理汉语的一条快捷方式。

[3] 奠定“语言文字系统工程”的基础

基于“两表”的“三化”乃至“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[35]，从而为进一步灵活多样的动态分析和计算机辅助处理奠定“语言文字系统工程”的基础[36]。

[4] 发现与记录

---

33 被揭开了“微妙关系”的神秘面纱。
34 意味着：可计算（或统计或分析）、易操作。
35 在“现实需要”与“理想目标”之间架设“桥梁”。
36 即有针对性地重用汉语语汇的“理想化认知模型”。

实际应用中，语言文字的形式与内容，通常总是联系在一起的。凡经过“语言文字系统工程”形式化处理的“音节”或“文本”序列[37]，在协同智能计算系统中，无论其形式信息还是其内容信息，都将一目了然[38]。用户个性化重用不过是该系统标准化重用的某些具体的组合变换[39]而已。通常会有一类例外，即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，本系统将自动记录该用户或该终端的原始输入信息，然后，与本系统长期协作的知识产权专家们一道共同对之进行复查和审核。

[5] 区分“字”与“词”的必要性

(1) 汉语的字与字组的关系（涉及字之间“接续”的问题）

区分“形字”与“音字”，是“汉语（间接）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”[40]的“切辞块”[41]与“断句读”[42]的困难如何解决（这是汉语内外教学和中文信息处理共同关注的问题）。

(2) 汉语与英语之间的结合（涉及与国际“接轨”的问题）

区分“音字”与“英词”，是“汉语（间接）形式化”的另一基本问题。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。如：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”（对自然人而言）之外，又增加了“分词”与“标注”的困难（对计算机而言）。

8.3.1.3 “五个限制”

通过“两表”限制“字与字组”总量，给出“序位编号”明确的类例，即形式消歧。

通过“音字”直接限制义项，给出“字的形式化定义”的解释方向，涉及迭交原理。

通过“数字”间接限制字组，给出“字组划分数字化”的计算方式，涉及等价原理。

通过“用字”直接限制义项，给出“义项呈现字组化”的线性组配，涉及组字公式。

通过“三注”间接限制义项，给出“义项呈现字组化”的立体选配，涉及内容消歧。

“两表”实际上是反映“两库”的两个一览总表。在形式上它们建立在数字计算机及其关系数据库和数据仓库的基础之上；在内容上它们是基于“相对完全归纳”的“语言事实”集合[43]。“两表”是标准化与个性化结合的汉语语汇理想化认知模型，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“字本位”结合的有益尝试。

8.3.2 字的定义形式化

“形式化”是符号逻辑、计算机软件、中文信息处理等多个领域通用的一个术语。基本含义就是符号化、结构化，目的是排除歧义（以利于计算机自动化处理）。

借助“两表”给出“字”的形式化定义（随之也就解决了“字组”的形式化定义）。

8.3.2.1 借助 GTCM 凸显：形字

从汉语文字学角度，可把属于视觉（形式）信息范畴的标准（形）字，如：1000 个最常用字，2500 个次常用字，7000 个通用字，含：国家标准 BG 2312 收纳的 6763 个通用字，…，统统归入

---

37 对汉语语汇而言就是充分利用 GTCM 的 4、5、6“进阶层式”和 GSCM 的 1、2、…、m“进阶层式”提供“音字序列”形式信息以及 GTCM 的 0，1，2，3，4“进阶层式”提供“形字序列”形式信息，优化中文信息处理的过程。

38 因为经过“两表”、“三化”乃至“三注”之后的汉语语汇知识信息数据处理系统可随时为用户重用和共享。

39 分与合——有针对性地的狭义处理（重构或重组）。

40 汉语“字本位”理论突出“字”在“汉语”中具有“基本结构单位”的地位。

41 汉语“字本位”理论突出基于“实字”的“辞”和在“实字”或“辞”的基础之上附加“虚字”的“块”。

42 由古代汉语延续下来徐教授的汉语“字本位”理论突出了“读”的“语气停顿”和“句”的“语义停顿”。

43 如《现代汉语词典》。

GTCM 第 4 进阶层式（“形式”类），即：导入该一览表，从而建立“**形字**”概念和“形式”类。

8.3.2.2 借助 GSCM 凸显：音字

从汉语语音学角度，可把属于听觉（形式）信息范畴的标准（音）字，如：1000 个最常用字，2500 个次常用字，7000 个通用字，含：国家标准 BG 2312 收纳的 6763 个通用字，…，统统归入 GSCM 第 1 进阶层式（“形式”类），即：导入该一览表，从而建立**音字**的概念和“形式”类。

8.3.2.3 “形字”与“音字”的关系

借助“两表”可展示：汉语“层面型结构”与“线串型结构”（见：图 3、图 4、图 5），凸显“形字”与“音字”。但在汉语实际应用中，“形字”与“音字”往往呈“迭交”状态。

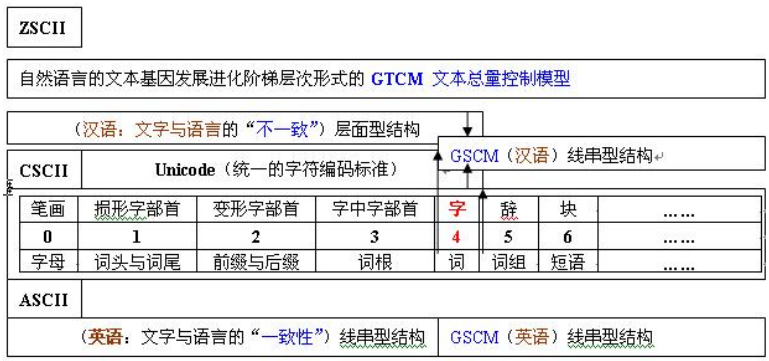
图 3 是“两表”中“字”的局部的示意图

由图 3 可见 GTCM 第 4 进阶层式与 GSCM 第 1 进阶层式，其表达对象是一致的。

8.3.2.4 （形式上的）“迭交”原理

在 GTCM 中，“形字”与“音字”是呈“迭交”状态[44]。在 GSCM 中，“音字”强调“字形与字音的一致性”，体现了汉语“形字”与“音字”在形式上“迭交”的特点。字音，却没有这种特点。字形与字音的说法，不能够揭示这种“迭交”关系。汉语“形声字”的“声符”（1325 个）具有“音字”的特点。可见没有“汉语拼音”时“音字”就存在。音译外来语的字也是“音字”。汉语“字形”的数目大大超出“字音”——汉语拼音的数目，因为，汉语拼音只有 400 多个音节，加上声调的变化，总共也不过 1000 多个。

图 4 是“两表”中“字”全局（同时涉及 GTCM 和 GSCM）的示意图。



由图 4 中只有“字”同时位于“层面型结构”与“线串型结构”的“迭交”处。汉语“形字”与“音字”的“迭交”关系在后面的实例 1 及图 10 和 11 中还有更具体的说明[45]。

简单地说，图 3 和图 4 对“迭交”关系的说明是这样的：首先，图 3 从局部突出说明“字”这个具有“迭交”特征的“形式”类，位于 GTCM 与 GSCM 的“枢纽”地位。接着，图 4 不仅强化“层面型结构”与“线串型结构”的“迭交”特征和“枢纽”地位，而且，展示了汉语与英语在结构（形式）上的一个根本区别，即：汉语的“两段性”[46]与英语的“连贯性”[47]。图中，与 ASCII 兼容的 ChSCII（中文标准信息交换码）是着重说明“字内信息”的处理机制。ASCII 和 ChSCII 都与 Unicode 兼容。ChSCII 是直接从 BIOS 进入汉语系统的基础代码[48]。

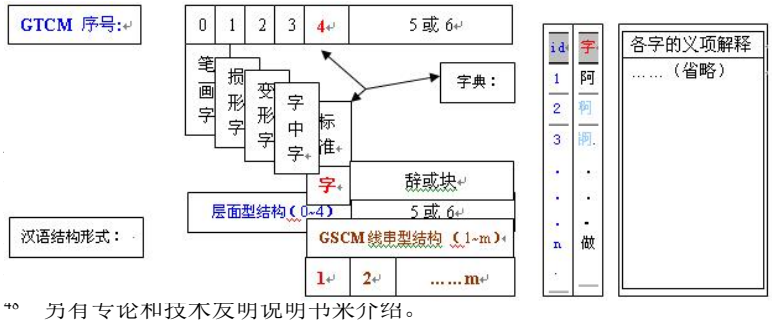


图 5 是字的形式化定义及其典型实施例示的意图。

在图 5 中，可见“形式”类

[46] 另有专论和技术反明说明书来介绍。

（理论上基于“两表”的抽象“类”定义）与“形式”例[49]，分别从理论与实际两方面体现了字的定义形式化和“迭交”原理（形式部分）及其应用实例[50]。读者可根据这个示例再造“表格化、数字化”电子字典，既能做到“数字”与“例字”同一并列，又能做到“形字”与“音字”同一并列。为“字与（各级）字组（形式）的间接计算乃至直接呈现”奠定“表格化、数字化”基础。

在图 5 中，笔者通过展示“字”的实例为编撰高质量的汉语数字化电子字典提供“直接呈现与间接计算的汉语形式化”例证。在理论上指明了“字”的“形式”类，同时，在实践上枚举了“字”的“形式”例。

综上所述，图 3、图 4、图 5 一起展示了：字，在汉语结构（形式）体系中的特殊地位，即：位于“层面型结构”与“线串型结构”的“迭交”处，具有“联结枢纽”的地位。体现了“形字”与“音字”同一并列的“迭交”原理。

8.3.2.5 语言基本结构单位“字”具有语音及其形态（与文字重叠）的形式特征

基于上述“迭交”原理及示例，有必要强调：作为语言基本结构单位的“字”[51]在形式方面表现出来的“形字”与“音字”的现象[52]值得重视。

语言学“形字”与文字学“字形”的关系。语言学“音字”与语音学“字音”的关系。语言学“实字”与语义学“字义”的关系。语言学“虚字”与语法学“字序”（或字位）的关系。语言学“用字”——微观语用学的“上下字”与“字解”的关系。

“字的定义形式化”涉及两个方面的“形式化”，即“字的形式”（即“音字”和“形字”）的“表格化、数字化”；“字的内容”[53]的“表格化、字组化”

“音字”和“形字”的区别：

首先，两者依据的表格参照系不同。其次，在狭义范围两者的形态一致。最后，必须指出：讨论“字与（各级）字组的（形式）关系”所依据的主要是 GSCM 中具有“语言特征”的“音字”的形式化定义。也就是说，在形式上我们可借助 GSCM 把“音形交融状态”的“字”从在 GTCM 中位于“层面型结构”与“线串型结构”的“迭交”状态分离出“音字”——位于“线串型结构”的“节点”（含“起点”）上。在内容上它可是具体的“实字”或“虚字”。

8.3.2.6 字与（各级）字组的（形式与内容）关系

只讨论“字与（各级）字组的（形式）关系”，至于“字与（各级）字组的（内容）关系”将在讨论“字的义项”的“直接呈现（组字基本公式）”与“间接呈现（‘三注’）”时再专门讨论。仅就形式而论，方块状“形字”与单音节“音字”（区别于仅仅表现字音的汉语拼音）是一对一的“迭交”关系。这体现了“音字”和“形字”的联系。

“字”的形式与内容好比是一个硬币的两个方面，当形式[54]确定之后，可进一步确定其内容[55]，反之亦然（意味着：形式与内容，可交替考虑，甚至结合考虑）。

字的含义的限定方式，涉及宏观与微观两个方面，前者，是从大环境上限定（如上下文及其所属领域）；后者，是由小环境来限定（如上下文的具体用字）。

字的定义形式化从形式上确定“字”的基本性质和总体范围，同时，也为形式上定义和划分“字组”确定了可区分[56]且可数数[57]的衡量尺度。

8.3.3 字组划分数字化

---

49 实践中基于“两表”的直观“例”——标准字典。
50 即：直观展示的形式化电子字典，其特点在于：编号 id“数字”与“例字”同一并列。
51 具有语音及其形态等形式特征。
52 后面还将分析：在内容方面表现出来的“实字”与“虚字”以及“用字”等现象。
53 即：义项，涉及：实字、虚字；用字、解字；…
54 好比确定了硬币一面，如“字”在“两表”中的具体位置——即所属的表和格；
55 好比确定硬币另一面，如“字”在“三注”中的具体领域——即所属具体学科。
56 基于字的字组直接呈现的基础。
57 基于字的字组间接计算的基础。

字与字组的划分（主要是字组细分）旨在为字与字组的间接计算与直接呈现奠定“表格化、数字化”转换基础。

8.3.3.1 展示汉语语汇的理想化认知模型是由理想模型转化而来的应用模型

基于 GSCM 的理想认知模型是在“线性代数方程组”（数学原理）和“计算机数据（仓）库的（一系列基础）表”（计算机科学原理与软件技术）的可实现范围以内的抽象认知模型（简称：基于 A 库的汉语认知模型）。它是由理想模型转化而来的应用模型，可直接展示字与（各级）字组（形式）及其相互关系。其中，汉语语汇部分的具体实现方式，见：图 6。

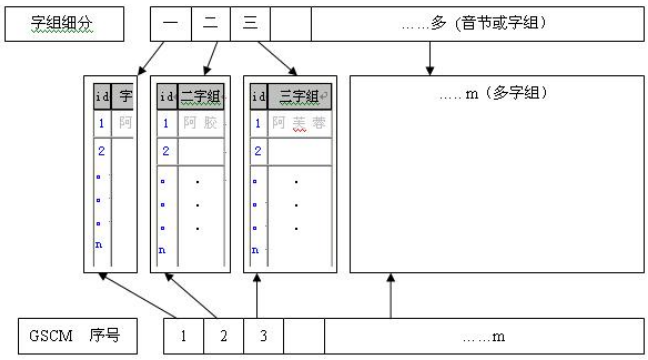


图 6 是字组划分数字化模型的示意图。

在图 6 中展示的是借助数学和汉语以及计算机数据库和数据仓库等工具而设计的一个理想认知模型的直观接口示意图。实际上，它体现了：基于 GSCM 的“字组数字化”原理及方法，即字与字组的直接展示与间接计算的原理及方法。与之配套的还有相关的原理及方法[58]。

在图 6 中蕴含认知科学、逻辑学、数学、计算机科学等领域的科学原理和技术方法。

[1] 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”或“模式识别”[59]。各种各样的“问题状况”及其“内部表示”涉及“知识的获取与表达”（局部理解）。“对（语言）关系”建立的“内部表示”涉及完整的“知识表达”（全局理解）。

例如：文本总量控制模型（见：图 1）和音节总量控制模型（见：图 2 以及图 6 的实施例）就是 Gene Culture 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”及其“内部表示”（动态模型）。这个模型及其实施例，是否被其它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评价或评估，不过，现在的实验和分析认为它具有可计算、可操作、可重用、可共享的特征[60]。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”乃至“三注”等具体“限制方式”，本系统可针对“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“电子字典”和各种标准化与个性化统一的“实用语汇工具”（含工具书）为汉语内外教学和中文信息处理，提供计算机辅助（CA）。

[2] 逻辑学原理

在图 6 所述“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(即：音字与音字符串)序列与“1，2，3，…，m”数字（或代码）序列之间“同义并列”。根据“同义并列，对应转换”公理[61]，任何两个形式信息体系，一旦“同义并列”，即可在各自的参照系中实行“对应转换”。

另外，“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。

[3] 数学原理

在图 6 所述“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“数字与文字”的“行”的序号（n）构成的矩阵序列，即“线性方程组”（$\Sigma a_{mn} x_n = b_m$），

58 因超范围，虽有助于拓宽思路，故不在此介绍。
59 对关系的识别。
60 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。
61 “序位逻辑”法则（即《理论融智学》的通则）。

其常数项，构成相应的矩阵序列。

[4] 计算机科学原理

在图 6 所述“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（即：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。

综上所述，图 6 所示的基于认知科学、逻辑学、数学和计算机科学的自然语言理解的理想化认知模型（形式体系）是可计算、可操作、可重用、可共享的[62]。在计算机中该理想认知模型与各自然人实际使用的个性化认知模型之间可构成一种高度协作且优势互补的协同智能计算关系。

例 1：在汉语“字本位”理论中展现的“字、辞、块、读、句”和“字、二字组、三字组、…、多字组”汉语认知模型（起初仅仅存在于徐通锵这个自然人的大脑之中）。

例 2：在工程融智学理论中展现的“0，1，2，3，4，5，6，7，8，9，10，11，12”（GTCM）和“1，2，3，…，m”(GSCM) 自然语言理解的理想认知模型（起初也仅存在于邹晓辉这个自然人的大脑之中，可是，一旦它们进入 Gene Culture 这台计算机成为一个形式体系之后就同时具备了无歧义地存在于任何一台计算机中的客观条件）。

例 3：“…,4，5，6，7，8，…”（GTCM）与“字、辞、块、读、句”之间构成的交集以及“1，2，3，…，m”(GSCM)或“一、三、…、多（字组）”（广义字组）与“字、二字组、三字组、…、多字组”（狭义字组）之间构成的交集，合成了汉语理解的理想认知模型（不仅可存在于徐通锵和邹晓辉这两个自然人的大脑之中，而且还可无歧义地存在于任何一台计算机系统之中，从而，也就更加方便其他人共享这一理想认知模型）。

例 4：根据“例 2”所述的自然语言理解的理想认知模型提供的原理和方法，把《现代汉语词典》导入“例 3”所述的汉语理解的理想认知模型之中，并适当地加以改造和重构，使之成为汉语语汇理解的理想实用模型（既便于计算机辅助汉语研究，又便于计算机辅助汉语教学，还便于进行各种各样的中文信息处理）。

8.3.4 义项呈现字组化

义项的字组化呈现，旨在：实现内容与形式之间、（协同智能计算系统的）前台与后台之间的相互转换。其原理有两条：一、“内涵减少，外延增加”；二、“同义并列，对应转换”。字的义项与用例字组或释义字组之间的关系，涉及 Gene Culture 微观语用学（如：限制义项范围的“用字”与被限制义项的“解字”之间的反变关系和互用关系，释义字组，基本组字公式，…）；释义字组[涉及：释义字组方阵和系列组字公式（其基础是基本组字公式，即：用字+解字 =二字释辞）与互用关系（如，在释义二字组中，用字与解字可对换互用——但每一次关注的“焦点”或“核心”是不同的）]，释义字组方阵的解析（如：1、仅仅由实字与实字构成的二字组方阵、三字组方阵、…多字组方阵；2、由实字与虚字构成的二字组方阵、三字组方阵、…多字组方阵；3、虚字与虚字构成的二字组方阵、…）与 Gene Culture 语义分析学[如：字的义项与字组的义项，直接释义法（基于组字公式与字组方阵，如：借助释义的字组、句子…）与间接释义法（基于同义并列，如：借助释义的字组、句子…；基于领域信息的划分和标注，如：借助“三注”的方法）]。

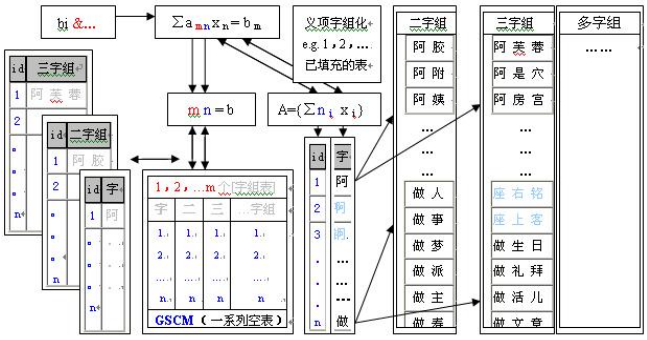
8.3.4.1 字的义项

“字的义项”的解释，可采用“释义的字组、句子、段落、篇章”等具体方式。本文仅采用“释义字组”这种基本形式。由于“义项”的解释或表达的形式，仅限于“字组呈现”的方式，所以，称为：义项呈现字组化。字的义项与释义字组之间可建立两种基本关系，由此获得：“字的义项呈现字组化”的两种方式；“释义字组”的形成或来源的两种基本途径。由于义项内容与

---

62 贯穿其中的“同义并列，等价代换”法则，即：形式信息处理的转换原理，简称：等价原理。

释义字组（形式）之间存在（“是或否”包含该字——其义项需要解释的字，简称：解字）两种可能，所以，这里假设：“是与否”两种情况，分别给出中文信息处理的相应策略及方法。当包含“解字”时，采取：直接呈现的策略及方法；当不包含“解字”时，采取间接呈现（标注）的策略及方法。



（计算机后台数字的分布函数与前台呈现的释义字组满足同义并列的条件）
基于 GSCM 的义项字典与字组用例，采用：“义项直接呈现字组化 ”的策略及方法。

[1] 义项的直接呈现

图 7 是“直接呈现”原理及实例示意图。直接呈现（采用音字线性组配）“释义字组”（由“线串型结构”包含的“字间信息”相互制约关系，实现义项收敛而达到解释或消歧的目的）——直接限制形式。依据“外延增加，内涵减少”的逻辑学原理，通过延长“释义字组”长度的方式，即：组配“释义字组”或“线串型结构”，达到限制被解释字与字组的“义项”的目的。

直接释义法，即：直接消歧法，其特点是：单刀直入——从形式方面直接逼近。由图 6 可见，义项用例就是释义字组——直接呈现具体的义项。这种线性延伸限制消歧是一种形式消歧法。

直接呈现“字的义项”的基本方式，即：以“解字”为“核心”，向左右两个方向直接“组字成语”的方式。其特征是：直接呈现“字的义项”的“释义字组”必须包含“解字”。根据逻辑学“内涵与外延的反变关系”和基础语言学“核心字与组字法”以及“两表”中体现的“字与字组的关系”，“直接呈现”，即：在“释义字组”中直接呈现“字的义项”的“收敛过程”。“两表”和“三化”使“此法”获得计算机辅助(突破仅依靠自然人的手工操作的局限，走上计算机自动化与自然人智能化之间分工协作的道路)，从而，得到系统地推广和普遍应用。“线性组配”，通过“左右限制”消除语义分歧，操作上是“形式组配”， 效果上是“内容消歧”。

（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）
附加“三注”的义项大典与用例大全，采用：“义项间接呈现（标注）字组化”的策略及方法。

[2] 义项的间接呈现

图 8 是“间接呈现（标注）”原理及实施例的示意图。

当“释义字组”不包含“解字”时，采用：间接呈现（标注）的策略及方法。间接呈现（基于一系列知识领域的立体标注）“释义字组”——间接限制形式。依据“同义并列”的原理和基于“两表”的“选域定向，测序定位”功能，通过计算机辅助分析“解字”及其“直接呈现”的“释义字组”或“线串型结构”这些形式信息，为“选域定向，测序定位”，搜寻其余的“释义字组”，达到限制“义项”的目的。这是：立体延伸限制消歧的方法，即：间接消歧法。特点是迂回包抄——从内容方面间接逼近。

间接呈现“字的义项”的基本方式，是基于“两表”的“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”）的方式。其特征是：间接呈现（标注）“解字”义项的“释义字组”（“标注字组”）不包含“解字”（或被释义的字组）。通过“领域限定”缩小“标注字组”查询范围（通过“三注”实现）。根据“等价原理”，比对“解字”义项（直接呈现“释义字组”）与“标注字组”之间是否存在“等价代换”的可能。本文仅限于字与（各级）字组的形式分析，所以，对“三注”不做具体深入的介绍。“间接呈现”，用“标注字组”来间接呈现“解字”的义项。基于“两表”和“三注”可使“此法”获得计算机辅助。“立体选

配”通过“领域限制”消除语义分歧，表面上是“内容查询”实质上还是“形式消歧”。

上述“两法”结合，即：“线性组配”与“立体选配”结合。“释义字组”与“标注字组”都是为限定“解字”的“义项”范围。“释义字组”或“标注字组”与“解字”关系，相当于义项的“例（用例）与类”的关系。前者，反映：语言事实（具体的“字组”形式——音节或文本这类“线串型结构”的组配形式）；后者，反映：认知概念［抽象的“字组”内容——脑中体现（指称对象或意象）的类例，通过“三注”可使之外化——即：在“两表”中得以体现］。

8.3.4.2 组字公式与字组方阵

汉语“字本位”理论，把语汇分为“字、辞、块”三种基本类型。在这里把“辞”与“块”统称为“语”。这样，“组字成语”的逆过程就是“分语为字”，其中，涉及：切“辞”、分“块”（即：从“语”中切分出“辞”与“块”）两个步骤。

字与（各级）字组的关系中，字与二字组的关系是基础，下面主要对“二字组”类型的“辞”与“块”做具体分析。如果把需解释其义项的“字”命名为“解字”，把限定“解字”义项范围的“字”命名为“用字”，那么限于“二字组”的“释义字组”就只有“释辞”与“释块”两种类型。

[1]“释辞”=“实字”+“实字”=“用字”+“解字”[63]。

[2]“释块”=“虚字”+“实字”=“用字”+“解字”；

[3]“释块”=“实字”+“虚字”=“用字”+“解字”；

[4]“释块”=“虚字”+“虚字”=“用字”+“解字”[64]。

上述[1]-[4]四个公式中的实字与虚字的关系，恰似一个方阵。故简称：字组方阵，也是：基本字组方阵。后续的“三字组”、“四字组”、…、“多字组”的“字组方阵”都可基于该“基本字组方阵”的原理而推衍出来。同理，后续“三字组”、“四字组”、…、“多字组”的“释辞公式”也都可基于该“基本释辞公式”的原理而推衍出来。同理，后续所有的**“组字公式”**都可基于该**“基本组字公式”**的原理而推衍出来。

例如：“释义三字组”也有“释辞”与“释块”两种类型。其中，

“释辞”=“实字”+“实辞”=“用字”+“解辞”；

“释辞”=“实辞”+“实字”=“用辞”+“解字”；

“释辞”=“实字”+“实字”+“实字”=“用字”+ “用字”+“解字”。

“释块”= …… 。进一步的划分或枚举，由于可类推（如“用块”、“解块”）和枚举（如“四字组”、“五字组”的“释义字组”），所以，本文均不再做具体介绍。

8.3.5 实例解析

8.3.5.1 实例 1：解析一个“字”——对汉语自身的结构分析

步骤：首先，解析“字内信息”，然后，解析“字间信息”。

［1］解析“字内信息”

基于笔画的（广义与狭义）“形字”，是“层面型结构”。对“层面型结构”的“形式化”处理是基于“GTCM 第 0，1，2，3，4 进阶层式”五个分表的“类”及“例”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 9 的一览总表）。

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	亠	乙	阿
3	丿	冂	冫	二	啊
4	丶	勹	口	十	锕
5	乙	冖	彡	厂	錒
6	…	…	…	…	…

图 9 是 GTCM 五个表的一览总表摘要介绍示意图。

由图 9 与图 1、图 4 和图 5 结合可说明：GTCM 第 0,1,2,3,4 进阶层式与计算机 FONTS

63 这是“基本释辞公式”，也是：基本组字公式。

64 “释链”=“虚字”+“虚字”=“用字”+“解字”。

（字库）以及 GBK 或 Unicode 的兼容关系。图 9 与图 10 结合可展示基于 ChSCII（中文标准信息交换码）的思路，实现 FONTS“形字”的“层面型结构”化“改造”——使“字内信息”成为可计算、可重用的“形式信息”。

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”又具有“可计算性”，因此从“形字”中提取必不可少的“字内信息”很方便。由于“形字”与“音字”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。

基于“音字”的“线串型结构”与基于“形字”的“层面型结构”是汉语的两大形式特点。

对“线串型结构”的“形式化”处理是基于“GSCM 第 1 进阶层式”这个分表的“类”及“例”而实现的。在计算机数据库和数据仓库中，表现为：由一组“数字（id）”与“音字（拼字音节的字符）”数据“同义并列”的一个“表”（见图 9 和图 2 的标准“字”表）。基于“音字”构成的（拼字）“字组”是通过“GSCM 第 2，…，m 进阶层式”的 m-1 个分表的“类”及“例”间接实现的。在计算机数据库和数据仓库中，表现为由 m-1 组“数字（id）”与“字组（拼字字符串）”数据“同义并列”的 m-1 个“表”（见图 2 的“字组”一览总表）。“汉语（间接）形式化”，主要立足于“音字”作为“线串型结构”的“节点”（含“起点”）可计算原理。

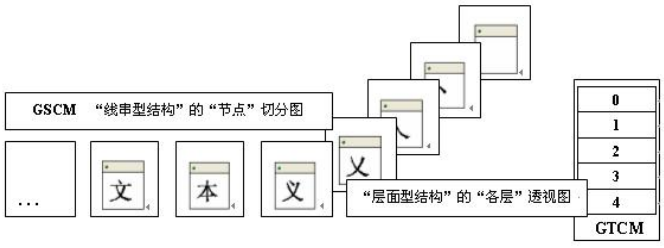


图 10 是形字与音字“迭交”原理及实例的示意图。

由图 10 可见“音字”切分为“节点”与“形字”拆分为“部件”。“层面型结构”顶层可透视音形“迭交”情形。如：图 10 中“义”这个“字”正位于“线串型结构”的“音字”与“层面型结构”的“形字”的“交汇处”。

[2] 解析“字间信息”

就“线串型结构”而论，图 10 中“文本”与“本义”虽然是两个可直接“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 10 中“义”这个“字”的“字内信息”涉及一个“乂”[65]和一个“点”[66]。其它字的“字内信息”解析与此同理。由此可见每一个“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，简称：汉语“字”的“迭交”原理。

“字形”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“拼音化”。表现出汉语总与拼音这根拐杖联系在一起的特点，又叫（现代汉语的）“一语双文”原理。

“音字”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。因“字间信息与字内信息”

---

65 字中字——结合图 5 和图 9 来理解。

66 笔画字——结合图 5 和图 9 来理解。

对“义项”都具有限制作用，故从“释义字组”的选取范围考虑，必须同时兼顾“音字”和“形字”两方面的语言信息（即：“字间信息与字内信息”）。

在图 10 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”是“迭交”关系。“音形字”的“声符”可视为“音字”的特例或原始类型。

在图 10 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”。

[3] 如何解析“文本义”这个“音字”串？

步骤 1：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”。步骤 2：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”。步骤 3：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”。步骤 4：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”。步骤 5：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”。以上“一字表、二字表、三字表”分别位于“GSCM 第 1，2，3 进阶层式”。同理，也可分析其它“线串型结构”。

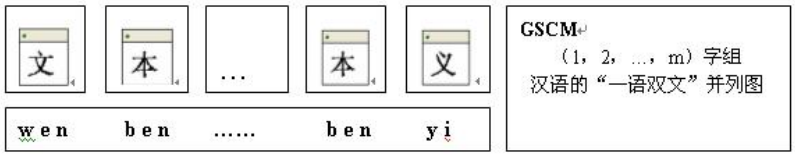


图 11 是一语双文示意图。由图 11 可见汉语具有（音）字与（由音字组配的拼字）字组和（由字母组配的拼音）音节和音节串两种记音符号。由此可见，汉语“音字”与汉语“拼音”（即“字音”的一种表现形式）并存的所谓“一语双文”现象。图 10 与图 11 结合可帮助我们更好地理解汉语语言学与“形字”相互“迭交”的“音字”[67]和汉语语音学的“字音”之间的区别与联系，见：在图 11 中上方的汉语“拼字”与下方的汉语“拼音”之间的“同义并列，对应转换”关系。

如果没有图 9 和图 10 对“义”字的拆分与叠合关系那种直观展现，如果没有图 11 对“一语双文”关系这种直观展现，那么，人们通常是不大容易理解“线串型结构”与“层面型结构”、“音字”与“形字”是如何构成具体的“迭交”关系的。

实例 1 和图 9、10、11 从汉语自身结构分析的角度展示了“分层组配”限制“层面型结构”的“释义”与“连串组配”限制“线串型结构”的“释义”的原理。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”。

8.3.5.2 实例 2：典型分析

[1] 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“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”[68]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

---

67 在图 10 中已通过“义”字的拆分与叠合的方式直观展现。

68 不过这里的“义项”已是对“本义”这个“字组”而言了。

性。…

[2] **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”，在它的前面只须增加一个“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对象“字形”的研究范围有关连。

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。

[3] **典型分析 3**：如把“典型分析 2”的研究向前推进一步，即：在“释辞”中扩大“字”这个解字的义项选择范围——扩大到“形字、音字、实字、虚字、用字”的选择范围。图 12 是“形、音、实、虚、用”关系的示意图。

无论你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在引入科学的语音学的同时实行了“汉语拼音方案”（之后就产生了所谓“一语双文”的情况）。（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，不仅古代汉语甚至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建立之后[69]，随着普通话的推广，汉语出现了“一语双文”[70]。于是，汉语在字形与字音（如：拼音）“分工”的过程中，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“形字与音字之间难以区分的不明确形式”。这样，在汉语普通话“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”。

[4] **典型分析 4**：“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。

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。

图 13 是“义项”与“释义字组”的（直接呈现）关系的示意图。

由图 13 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

69 这也是“汉语拼音方案”推行的结果。

70 如考虑汉英乃至汉外语言，再加上汉语方言，那么笔者甚至可断言：汉语是“双文双语”乃至“双文多语”。

方法（“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。其它可类推的部分省略。

[5] 典型分析 5：“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。

图 14 是“字的义项”与“释义字组”及“标注字组”（直接和间接呈现）关系示意图。

由图 14 可见：通过汉语直接呈现的“义”这个字的义项的“用例”不仅与英语的对应词语之间可实现双语同义并列（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“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”（“行列限制法”），围绕“（领域知识）参照系”展开。

8.3.6 结语

本文主要是从“中文信息处理”实际工程的需要和应用的角度，探讨“字与字组的关系”。

在进一步整理“两表”、“三化”及“三注”具体研究过程中，笔者归纳出了“基本组字公式”、“组字公式”和“字组方阵”。与“组字法”和“核心字”有“异曲同工”或“遥相呼应”的效果。这也许会延续汉语“字本位”理论的一个重要思路。

本文的真实意图是：为“开发基于汉语且兼容英语的高性能计算机的输出输入系统（BIOS）和操作系统（OS）及协同智能计算系统（Man-Com-Net）”而探索“汉语（间接）形式化”新路。在有目标的研究与有意识的应用中，笔者发现汉语“字本位”理论虽有争议但更有启示和挑战。

例证 1：基于“字内信息”处理，笔者提出以“ChSCII”为基础，与“ASCII”和“UNICODE”兼容的“ChBIOS”和“ChFONTS”设计方案。

例证 2：基于“字间信息”处理，笔者提出以“组字公式”和“字组方阵”为基础，以“字”的“迭交”原理为枢纽，“形、音、实、虚、用”一体化限制“解字”计算机辅助研究与教学系统的设计方案。这有利于（基于间接形式化的）双文双语信息处理和操作系统双文双语界面的开发。

例证 3：基于“字外信息”处理，笔者提出以“文字、语音、语义、语法、语用”等“语言文字信息标注”为基础，不断增加“通用常识信息标注”和“专用知识信息标注”规模的“组字成语-组字成句-组语成句-组句成段-组段成篇”计算机辅助研究与教学系统的设计方案。这有利于（基于“生产式教学”和“一体化管理”的）双文双语工具和多科知识创新环境软件的开发。

以上设计思路采纳了改进或优化之后的汉语字本位理论的研究成果，融合了语言学和信息学（含：计算语言学）成果，它们有着十分广阔的应用前景，希望能引起更多学者的关注与研究。

参考文献

David M. Kroenke[美] DATABASE PROCESSING——Fundamental，Design & Implementation （Seventh Edition）。北京大学计算语言学研究所，2000，《计算语言学文集》（第 4 集）1-254 页。

北京大学数学力学系几何与代数教研室代数小组 1978《高等代数》人民教育出版社 1-49，102-149，376-398 页。
陈肇雄主编，1992，《机器翻译研究进展》，电子工业出版社，1-564 页。
方　立，1993，《美国理论语言学研究》，北京语言学院出版社，1-240 页。
冯志伟，2002，《发挥汉语拼音在信息时代的作用》，载《 语文现代化论文集》，商务印书馆，41-44 页。
黄河燕主编，2002，《机器翻译研究进展》，电子工业出版社，1-282 页。
黄增阳，1998，《HNC（概念层次网络）理论——计算机理解自然语言的新思路》，清华大学出版社。
康博创作室 2001《SQL Server 2000 数据仓库设计和使用指南》清华大学出版社，14-36，49-69，113-230 页。
刘叔新，1996，《词语强制搭配的语义关系类别及其性质》，载《语言学论辑》，北京语言学院出版社，1-17 页。
鲁　川，2001，《汉语语法的意合网络》，商务印书馆。
施伯乐等译，2001，《数据库处理——基础、设计与实现》，电子工业出版社。
苏培成等，2002，《语文现代化论文集》，商务印书馆。
石　锋，1995，《汉语研究在海外》，北京语言学院出版社。
熊全淹，1978，《近世代数》，上海科学技术出版社。
徐通锵，2001，《基础语言学教程》，北京大学出版社，19-36 页，178-237 页。
徐通锵，1997，《语言论》，东北师范大学出版社， 295-442 页。
喻云根 1994，《英汉对比语言学》，北京工业大学出版社， 69-99 页。
张学文，2003，《组成论》，中国科学技术大学出版社， 44-56 页，246-252 页。
张志公，1996，《汉语简论》，载《汉语辞章学论集》，人民教育出版社。
朱志凯，1995，《逻辑与方法》，人民出版社，3-32，225-287，229-304 页。
邹晓辉，2004，《论影响人类未来的五大系统工程之间的关系》，《熵.信息.复杂性》第 86 期。
邹晓辉，2002，《协同智能计算语言数据库的设计方法》，《潜科学》第 32 期。
邹晓辉，2004，《义项语汇典例（SVDE）的总量控制模型》，《第五届(国际)汉语词汇语义学研讨会论文集》。
邹晓辉，2004，《字的形式化定义》，《潜科学》第 38 期。
邹晓辉，2004，《字组的划分方法》，《潜科学》第 38 期。
邹晓辉，2004，《字与字组的关系》，《潜科学》第 39 期。
邹晓辉，2004，《重构“概念分类体系”的新思路与新方法》，《潜科学》第 40 期。
邹晓辉，2004，《优化“语义信息处理”的新方法与实施例》，《潜科学》第 40 期。
邹晓辉，2004，《解析“字与字组的关系”探索“汉语形式化”新路》，《潜科学》第 41 期。

第六届（国际）汉语词汇语义学研讨会（论文）

重构“概念分类体系”的新思路与新方法

——从“语义三角”到“语法关系”再到“语义三棱”

摘要：从“语义三角”到“语法关系”再到“语义三棱”，以一个新视角看词汇语义学的概念分类体系。

关键词：语义三角、语法关系、语义三棱、概念分类体系

一、绪言

本文拟在汉语词汇语义学基本理论领域就“概念分类体系”作一些新的探讨。特点：从最基本的概念——“范畴”入手，以一个新视角看“概念分类体系”。重点：揭示“意”与“义”之间的“短程线”，从“语义三角”到“语法关系”再到“语义三棱”通畅思路。研究途径：通过一个简单的几何模型，显示“短程线”，借助“语义三棱”，探讨“基本范畴体系”统帅和驾驭“概念分类体系”的新思路与新方法。对“短程线”及“语义三棱”的探讨和应用，限于：几何

分析及协同智能计算语言数据库及知识库可操作范围。基本假设：“语义三棱”表达的“基本范畴框架”是统帅和驾驭“关系数据库”表达的“概念分类体系”的简捷方式。知识贡献：揭示“语义三棱”的实质，明确提出：基于“四大范畴框架”重构“概念分类体系”的新思路与新方法。

二、综述

本文对“基本范畴”和“范畴体系”的探讨，思想上的渊源：1、可追溯到中国远古时期《周易》的“形而上为用，形而下为器”以及先秦时期《老子》的“道”；2、可追溯到古希腊时期柏拉图《理想国》的“理念论”以及亚里士多德《工具论》的“范畴篇”。哲学上，贯通“实体论”、“认识论”、“语言哲学”和“信息哲学”四个认知阶段。由于主题、领域以及篇幅等限制，本文对上述思想不做具体介绍和分析，仅指出其离散分布这个事实。下面我们从 G. Frege 开始，探讨的重点和关注的焦点：从“语义三角”到“语法关系”再到“语义三棱”贯通其思路涉及“物、意、文、义”四个范畴。

本文指出：虽然 C.K.Ogden 和 I.A.Richards1923 年出版的《意义之意义》(The Meaning of Meaning) 明确提出了“语义三角”（semantic triangle：thing，thought，word；或 referent，concept，symbol）的说法，同时，还给出了直观示意图。但是，1892 年 G. Frege 发表的《意义与所指》(über Sinn und Bedeutung)明确区分“词所表达的意义和词所指的事物”的事实，早已揭示出（词所指的）事物、（词所表达的）意义、词语（文字的形 / 语言的音 / 符号表达式）三者之间的“语义三角”。为便于比较、借鉴和思考，让我们来回顾一下“语义三角”及其提出所产生的正负两方面的影响。

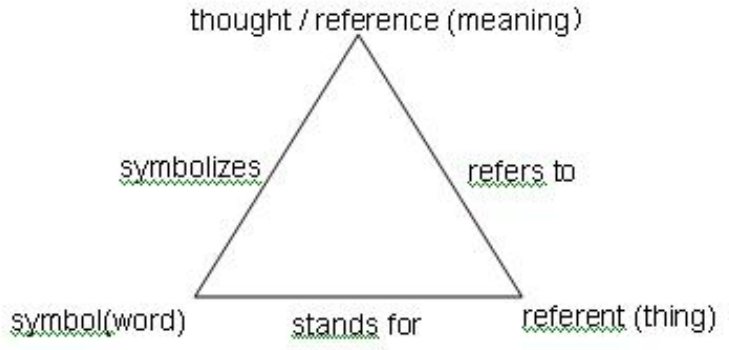


图 1

图 1 是 1923 年在《意义之意义》(The Meaning of Meaning)这部专著中 C.K.Ogden 和 I.A.Richards 给出的“语义三角”示意图。其中，括号里的“thing（物），meaning（意义），（文）word”提示本文关注的焦点，可视为基于词汇语义探讨而限制或明确的词语或概念，因为，“referent（所指），thought（思想），symbol（符号）”的说法过于含混而不便于以下精细分析和讨论的连贯性。“语义三角”的确有“简化”认识与表达的作用，因而有利于推广普及。例 1：1996 年在《语言与现代逻辑》这部专著中周斌武、张国梁介绍了“语义三角”（引文涉及：Montague 和赵元任等），还给出多种形式的“语义三角”变式图（本文省略）。

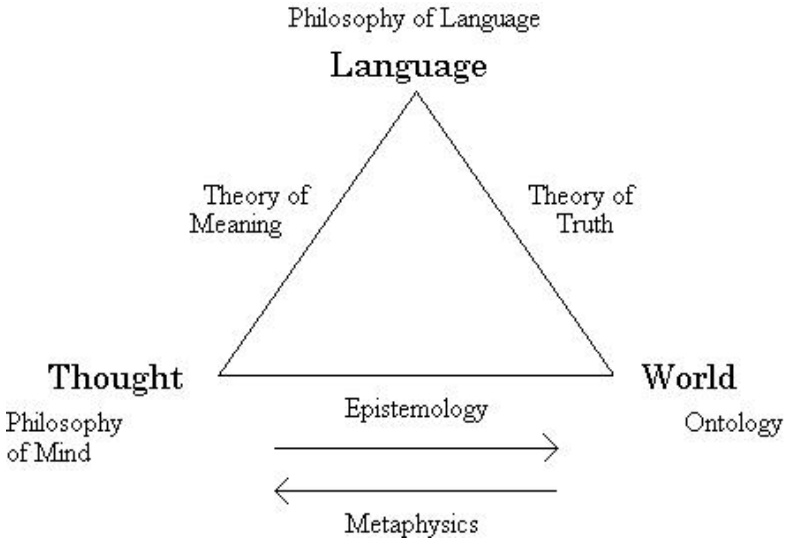


图 2

图 2 是 1999 年在《Meaning and Truth》“(i) Analytic philosophy and the philosophy of language”中 Mike Beaney 给出 the philosophical triangle：world，thought，language 示意图（即：例 2）。

以上是“语义三角”的正面影响。例 3：2003 年在《关于“语义三角”之我见》中李锡胤指出“语义三角”把“符号——概念——所指物”三者放在一个平面上，容易引起误解。这是“语义三角”的负面影响。

另一个值得注意的事实是：1915 年出版的《普通语言学教程》(Cours de linguistique générale)中，F.Saussure 在“语义三角”提出之后曾经进一步揭示出“语法关系”的重要性。用 F.Saussure 的话说，就是：语言好像下棋。如：在下棋时如果缺一个“车”，则可用一个小木块代替。此时，有无“车”字（此处把例子由国际象棋改为中国象棋）无关紧要，只要按“车”的“走法”去“走”就行。也就是说：只要按“车”和其它“棋子”的“关系”去“走”，“车”的实体性质或价值也就会体现出来。由此悟出一个道理：语言里每一个单位（如：词）看来是实体，其实不是。因此，F.Saussure 符号学理论认为：孤立的语言单位不是实体，有资格称为实体的东西来自价值，价值来自关系，关系构成系统。语言的诸单位，只有在语言系统里活动才有价值。

如果从 1892 年发表的《意义与所指》（G. Frege）开始计算，那么，语言哲学的意义理论，揭示出：事物、意义、词语（文）三者的关系，至今已过去 113 年。如果从 1915 年出版的《普通语言学教程》（F.Saussure）开始计算，那么，普通语言学的语法理论，揭示出：语言的诸单位，只有在语言系统里活动才有价值，至今已过去 90 年。即使从 1923 年出版的《意义之意义》(The Meaning of Meaning)这部专著（C.K.Ogden 和 I.A.Richards）开始计算，至今也已过去 82 年。C.K.Ogden 和 I.A.Richards 给出“语义三角”模型，虽然突出了“G. Frege 强调的（词的）意义与所指”，但却“屏蔽”掉了通向“F.Saussure 强调的（语言诸单位的）关系”的“短程线”。原本可在 20 世纪初期融合的两项深刻理论，就这样失之交臂（近一个世纪）。于是，对“语义三角”与“语法关系”（“语义三棱”将显示：实际上涉及的“关系”远不止于“语法关系”）的“本质联系”的发现，不得不经历相当漫长的岁月。与“语义三棱”相比较：“语义三角”的“平面视角”的确不利于人们把 G. Frege 与 F.Saussure 各自独特的见解联系到一起，对学界发现“短程线”或进一步认知“语义三棱”的确有“屏障”作用（例 3 也是一个佐证）。

笔者在 1976 至 1980 年期间开始尝试探索“语言与知识的计量和信息与智能的本质等前瞻性

问题”，1994 至 2000 年期间开始尝试系统应用和整理自己长期探索的认识成果并撰写了《中国企业知识产权战略（系列专栏文章）》、《一种知识信息数据处理方法及产品（发明）》和《融智学新范式》，明确地提出“义、文、物、意”融智概念体系（通论）、信息处理法则（通则）和多元数表达式（通式），从而公开了融智学理论框架。笔者虽然接触过 G. Frege 与 F. Saussure 的观点，但是，从未自觉地把两人的见解联系到一起。2000 至 2005 初，随着融智学思想体系（理论、工程和应用三个部分）的逐步完善，笔者日益体会到“语义三棱”的功用巨大，于是回过头来分析 G. Frege 与 F. Saussure 各自的独到见解。这时，才突然意识到“这对学术界来说可能是一个迟到的顿悟”：G. Frege 对“语义”的认识（着重于语言哲学的意义理论——涉及：数学、逻辑、语言、语义的探讨）与 F. Saussure 对“语法”的认识（着重于普通语言学——涉及：语言实践、语言系统、语法理论、语法实践的探讨）之间，存在一条无形的长期被“语义三角”的平面视角“屏蔽”而未被发现的“短程线”（可直接连通“语义三角”与“语义三棱”）。

狭义融智学作为一门研究自然人与计算机之间如何实现高度协作且优势互补的学问，在“人类智能”与“人工智能”之后提出了“协同智能”及其“融智概念体系”（其几何模型，可称之为：“语义三棱”）。从基础语言学和计算语言学的角度（即：限定在语言学的范围）来看，可把“语义三棱”视为整合“语义三角”与“语法关系”的理论模型，也可称之为：“融智三棱”。本文在汉语词汇语义层面讨论：“语义三棱”模型，对重构“概念分类体系”，可能带来哪些新思路和新方法。

三、方法

“迟到的顿悟”告诉我们：从理论上发现这条无形的“短程线”相当不容易。可是，一旦基础理论突破之后，再回过头来概括地描述它，则相对容易得多。如何表述这条无形的“短程线”才能做到“深入浅出、简明扼要、提纲挈领、恰到好处”呢？经过无数次尝试，我们终于想到了一个大胆的假设，即：借助一个简单的几何模型，来说明“意”与“义”之间的“短程线”这个深奥的理论发现。

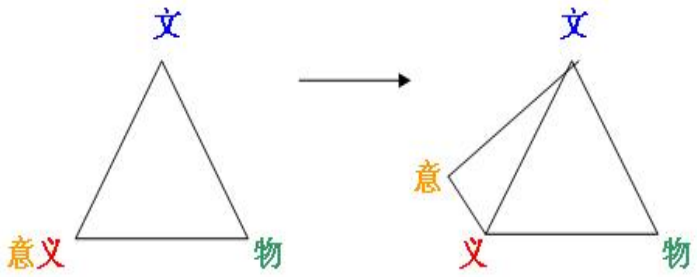


图 3

图 3 是“短程线”变化示意图。

下面用程序化语言与附图相结合的方式来说明“意”与“义”之间的“短程线”是如何变化的——涉及这个发现的“顿悟”过程及其逆过程。第一步，统一名称：1、给出“语义三角”三个顶点的名称：物、文、意义；2、给出“语义三棱”四个顶点的名称：物、文、意、义。第二步，画出直观的几何图形：一个“三角形”和一个“三棱体”。第三步，分别在“三角形”和“三棱体”的各个顶点填写相应的名称：1、把“物、意义、文”三个顶点的名称，分别填写到**图 3**“三角形”的三个顶点处——由此获得便于对比的“语义三角”模型图；2、把“物、文、意、义”四个顶点的名称，分别写到**图 3**“三棱体”的四个顶点处——由此获得“语义三棱”模型图。第四步，比较“两个模型”的区别和联系：1、联系：两模型图的“物、文”两对“顶点”彼此相同；

2、区别：“意义”占据“语义三角”的一个顶点，“意、义”占据“语义三棱”的两个顶点。这是静态方式观察的结果。接着，用动态方式观察：如果把**图 3** 的“三角形”视为“三棱体”的一个面，那么，只要旋转一定的角度，就能获得**图 3** 的“三棱体”的视角。这样，就能发现：在“三棱体”中“泾渭分明”的两点（“意”和“义”之间的“短程线”显现出来），在“三角形”中“迭合”成了一点（“意”和“义”之间的“短程线”被“意义”隐含）。可见，思维方式与观察角度之间息息相关。视角不同，所见不同。所见不同，想法不同，反之亦然。思维方式与视角不同，想到与看到的东西也就不一样。

再进一步，我们一起来做一个理想实验.。设想 1：当距离足够远时，可发现“四点合为一点”的情形。这是“置身事外”的情形。设想 2：当我们进入到“三棱体”之中的时候，能真真切切地感受到“四点完全分离”的情形——这是“身临其境”的情形。我们还可以做一个实际实验. 实验 1：绕着一个大大的“三棱体”周围转，可观察到“四点分离”的各个局部。这是“盲人摸象”的情形。实验 2：用纸做一个“三棱体”放在手中旋转，甚至折叠变换。可观察到其它不同的情形，如：“四点中有三点合为一点”可观察到一条“线段”和“四点中有两点合为一点”可观察到一个“三角形”。这是在现实与理想之间过渡的情形。从中可见：“三棱体四点变换过程中”蕴涵的“分与合”机制。当这“四点”是“物、意、文、义”的时候，我们可看到世界的千变万化的“分与合”机理。“语义三棱”揭示“短程线”的途径：一旦“义、文、物、意”范畴体系形成，即可发现“意”和“义”既可“一分为二”也可“合而为一”（即：“意义”）。稍微再进一步，就可发现“意”与“义”之间那段可改变的“距离”，恰似一条无形的“短程线”。通常情况下，它是被“意义”紧紧连在一起的（相当于：被“意义”这张无形的网“屏蔽”了）。至此，用一个十分简单的几何模型，就使一个“非显而易见”的关系“一目了然”——变得“显而易见”。“语义三角”与“语义三棱”的区别和联系，由此也可直观领悟一二。

通行的观点认为“意义”是一个“（双字）词”，其对应的“英语单词”是“meaning”，基于“词是不可分的最小的语言单位”的观点，“意义”不可分，被视为很正常，学界也习惯于把“意与义”两个“字”视为两个“语素”而不是两个“（独字）词”——至少不认为三个概念具有“短程线”揭示的相互关系。“意义”可分解成“意”与“义”的必要性和可能性，自然也就“视而不见”。认真比较一下“语义三棱”与“语义三角”，不难发现：仅从“词”（word）的观点（即：“词本位”的观点）来看，似乎什么问题也不会发现。因为，“事物（thing）、意义（meaning）”对应地来看，似乎都不能再“切分”了。但是，如果从“字”的观点（即：“字本位”的观点）来看，那么，“事物、意义”都能再分。其中，实质性的差别主要涉及：对“意义”的理解。理论融智学，正是从“意义”这个神秘的“屏障”中揭示出其中所隐含的“短程线”这个奥妙而实现由“语义三角”到“语义三棱”的认识飞跃的。

四、结果与结论

“语义三棱”的发现，揭示出：一个长期未曾被学界注意的事实——还有比“意义”更基本的概念“意”与“义”。认识上进这一步非同寻常。它不仅揭示出“概念体系”和“概念分类体系”建立的根基是否牢固的问题，而且，还揭示出一系列与之密切相关的重大认知问题。领悟到这些，笔者心里自然为之一振。毕竟二十几年的探索，现在终于揭示出一个根本性理论问题的实质。不过，为慎重起见，笔者仅仅在小范围内谈论这个发现的过程及其背后的原因，把主要时间和精力用于继续揭示其中蕴涵的一系列重要原理及其可能的重大作用。如：通过融智学系列文章、讲座和学术对话（与部分相关领域的专家之间的通话或通信）的方式逐步传播、交流和探讨。实践和实验证明：笔者 2000 发表的《融智学新范式》和《一种知识信息数据处理方法及产品》的确公开了一个重要理论的基本框架——“语义三棱”。

随着较为全面的检索查询以及进一步地学习、探索和国际国内学术交流，特别是经过与系统

科学、语言学、数学、计算机科学、计算语言学、认知心理学、人工智能乃至语言哲学和信息哲学等多个相关领域学术前沿的有关学科专家之间的交流，2005《潜科学》第 39 期（学术期刊）18“融智学专著及其知识要点和基本术语”（二）栏目以“融智学与以往的知识学问之间的渊源关系”为题目发表了直观描述“物质世界、思想（精神世界）、语言（符号世界）、关系（抽象的序位世界）”的“语义三棱”模型。

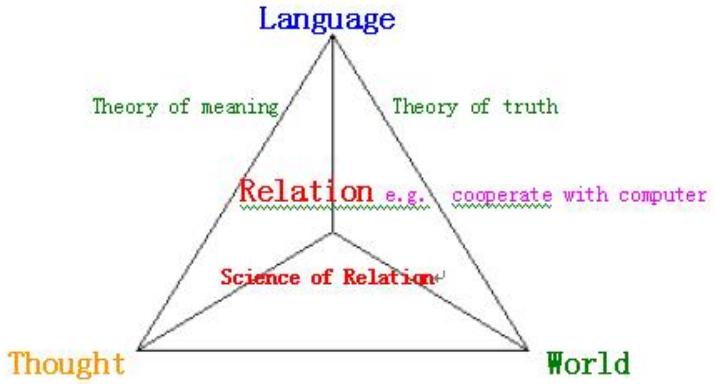


图 4

图 4 是“语义三棱”模型示意图。

由图 4 可见：“（对象）世界、思想（概念）、语言（符号）、关系（序位）”，简称：“物、意、文、义”四大基本范畴。括号内的说明表示限定在词汇语义层面的“语义三棱”模型，可借助计算机关系数据库进行描述。这不是一个简单的发现，既不是几何游戏，也不是文字游戏，而是发现了一个大原理。它蕴涵着：非常深刻的内容和丰富多彩的形式。一旦人们普遍理解了“语义三棱”模型，就可利用它的各个顶点的关系直观地分析和展现各种深刻的道理。其中也包含周斌武、张国梁看到的变换与李锡胤指出的关系，内容远远不止于此。不过，本文仅从基础语言学和计算语言学的角度来看，尽量限定在汉语词汇语义学基本理论的“概念分类体系”的范围，“语义三角”与“语法关系”如何整合在一起，从而，得到汉语词汇语义层面的“语义三棱”模型。

被“屏蔽”了近一个世纪的那条无形的“短程线”，至此终于被用汉语思维的中国人发现，并被明确无误地展现了出来。这是一个好的开端。我们知道，自从《意义与所指》与《普通语言学教程》发表以来，围绕着：意义的承载单位究竟是“词、短语、句子、…，还是整个语言系统”，产生了一系列不同的意义理论或语义理论与语法理论，却无一从“语义三角”与“语法关系”的联系来考察“语义三棱”。这个问题的重要性，不在于解释其原理的几何变换，而在于这种变换揭示出来的思维方式的变换。换一句话说，“语义三棱”比“语义三角”与“语法关系”可揭示：更深刻的内容和更丰富的形式以及更多更重要的科学原理。“语义三棱”甚至可揭示出以下过去未曾被系统认知的领域（含：本文的关注点多处）：1、把“relation（关系）”（含：“语法关系”）范畴从“语义三角”中独立出来。见：图 3-6。2、确立“四大基本范畴”具有从根基上优化“概念分类体系”的作用。3、发现并概括“八大关系”是“关系”的“基本分类体系”，也是新的“概念分类体系”的一个重要方面。为建立：序位模型，使“关系”本质的阐述在“认知图式”上获得新的统一。4、把对“意义”的“认知”推进到“意”与“义”的细分阶段。其重要性已明确。5、把对“事物”的“认知”推进到“物”与“事”的细分阶段。“可拓学”是证实“物”与“事”区分的重要性的一个佐证。6、概括并建构“八大形式”。这是新的“概念分类体系”的另一个重要方面。7、揭示“信息”现象及其概念的深层本质（定义）以及更为基础的划分（分类）方法。这是信息科学一直关注的重要的问题。8、揭示“语法”与“语义”的区别和联系的深层机理。这是理论语言学和计算语言学一直关注的问题。9、优化：“世界观”和“方法论”。这是哲学方法论十分关注的问题。10、优化“概念分类体系”……

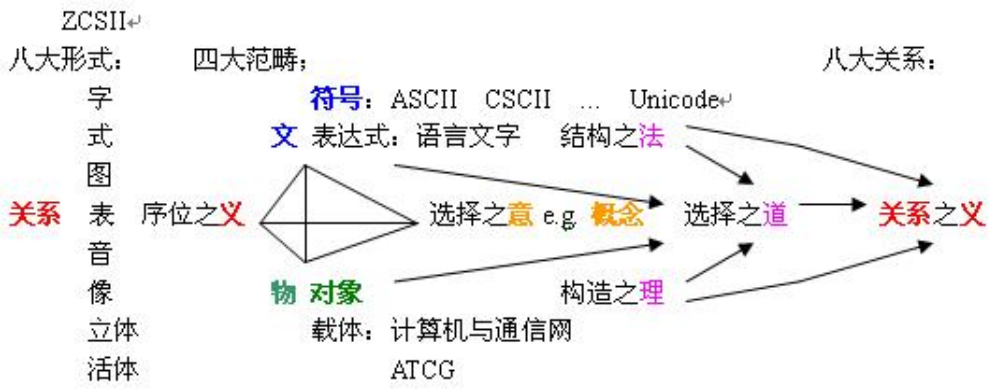


图 5

图 5 是“语义三棱”的四大范畴统帅和驾驭“八大形式”与“八大关系”总体概况示意图。

由图 5 可见：表现为“八大形式”的“文”与“物”两个范畴，被（表现为“八大关系”的）“义”结合在一起。“意”，表现为“协同智能主体”的一系列“选择”，虽然可洞察“八大关系”，但必须由“八大形式”体现。所谓“概念”和“概念体系”乃至“概念分类体系”只不过是我们这些“协同智能主体”的“成员们”所做出的一系列“选择”——记录在人脑中就是知识信息、记录在电脑中就是数据信息，如此而已。“概念分类体系”是“八大形式”体现的“八大关系”的派生子系统。现在，我们已清楚：“物（对象系统）、文（符号系统）”是可把握的——如：重用或再现、共享或分享，而“意（概念系统）、义（关系体系）”一旦脱离开“物、文”则显得虚无缥缈。科学技术和文化艺术的成就无一不是建立在“物、文”之上。一旦明白这个道理，新的“概念分类体系”的总体框架，就清楚了。基于“八大形式”来建立或展示“概念分类体系”，不仅包括：对“八大关系”的认识或理解（含：自然的奥妙），而且，还包括：“知、情、意、行、个性”这些与心理因素（含：心理的奥妙）。由此可见，基于“语义三棱”，可构建一个新的“概念分类体系”。总体框架：由“物、文、义”三大系统作为我们做出一系列“选择”的基础，由“文”记录这一系列“选择”——“意” 就是新的“概念体系”，对之实施标准化管理即构成新的“概念分类体系”。具体操作必须形式化——符号化与表格化。关键就看：标准化与个性化之间如何统一?下面看与词汇语义探讨密切相关的语言学各主干学科领域在“语义三棱”中的位置。

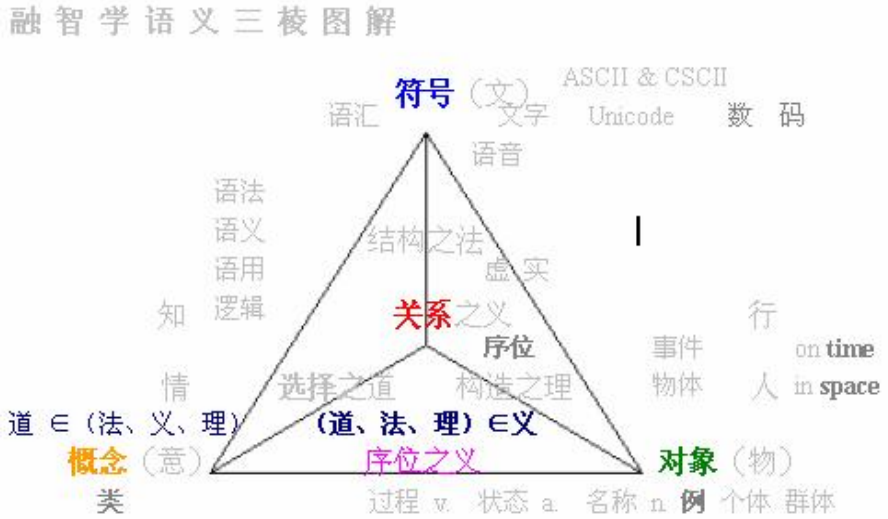


图 6

图 6 是语言学与“语义三棱”示意图。

由图 6 可见：语言学的“文字、语音、语义、语法、语用”等分支学科构成的静态理论体系。“词汇”是如何“在语言系统里活动”的呢？“语义三棱”可为人们提供什么样的有益思路呢？

借助位于“语义三棱”四个顶点的“物（对象）、意（概念）、文（符号）、义（关系）”的各种变换，人们可有各种相应的思路变换。例如：借助基本范畴框架，一方面，可说明：现象（物、意、文）与本质（义）两方面；另一方面，可说明：“概念分类体系”（意）与“对象分类体系”、“关系分类体系”、“符号分类体系”（物、义、文）两方面；再一方面，还可说明：“符号分类体系”（文）与“对象分类体系”、“概念分类体系”、“关系分类体系”（物、意、义）两方面；… 。对本文而言，“词”（文）与“概念”（意），是最重要的一对关系。如何依托“词汇分类体系”（文）实际把握“概念分类体系”（意）的各种变换？这是最关键的问题。 静态的理论体系与动态的实践系统，都可在一个好的“概念分类体系”（意）中把握，进而在一个好的“词汇分类体系”（文）中体现，反之亦然。借助“语义三棱”各种变换，可加速优选这两个体系的进程。既然已认识到这种程度了，那么，具体构建“概念分类体系”（意）的过程也就转变成：具体构建“词汇分类体系”（文）的过程。限定在汉语词汇领域，就是：以计算机数据库和数据仓库的方式，构建能正确体现“字与（各级）字组的关系”的词汇系统。其根基就是“字与二字组的关系”，中文信息处理就是“字内信息、字间信息、字外信息”的处理。于是，就有一个可操作的好方案：从宏观上以“基本范畴”统帅“概念体系”（意）；从微观上用“字组公式及字组阵列”驾驭“词汇体系”（文）——易于重用并共享的系统（见：《优化“语义信息处理”的新方法及实施例》）。

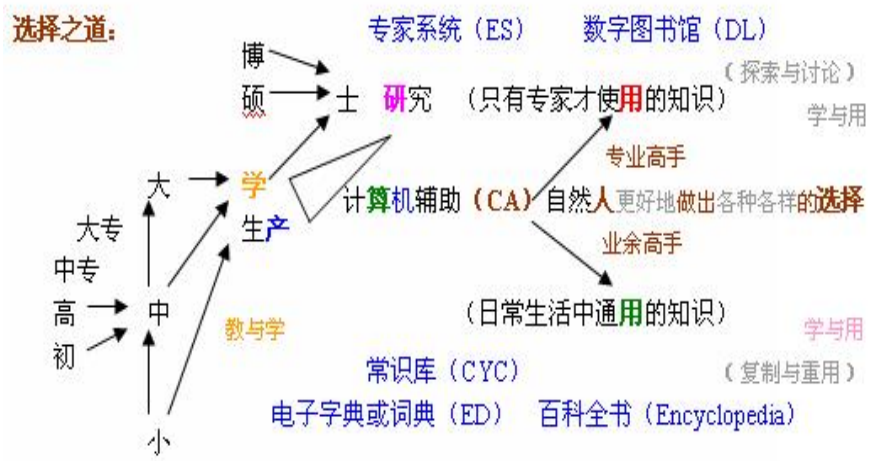


图 7

图 7 是各级知识重用系统关系图。

由图 7 可见，各种实际的计算机辅助（CA）应用系统，如：ED，CYC，ES，DL 等构成的“知识重用体系”，不仅与“知识等级划分”之间优化互动，而且，还与“产、学、研、用、算”五方面优化互动，的，其中，前四方面自然人是主体，后一方面计算机是代理。新的“概念及词汇分类系统”就是在上述各级各系列“中文信息处理”应用系统之间优化互动过程中不断优化的，其间“语义三棱”均起着“简化”或“条理化”的功用。

五、结语

综上所述，不仅“意义”这个概念再分解成为“（主观的）意”与“（客观的）义”两个范畴具有重要作用，而且，基于汉语特点建立的“语义三棱”——“物、意、文、义”对提出重构“概念分类体系”的新思路和新方法，也有宏观指导作用。

就汉语特点而论，与英语比较，两者最大区别是“字”与“词（word）”的不同。由于“字”与“词（word）”都属于“文”的范畴，根据“语义三棱”原理，两者都必然与“物、意、义”之间发生关系，因此，从“范畴”和“概念”来看，两者之间实际上又是“和而不同”。“和”是因为两者作为各自“语言系统”中最基本的单位“符号”均可被用于表达“概念、对象、关系”。“不同”主要在于：“形式及形态的识别”与“内容及知识的表达”两方面的差异。“字”与“词”的“不同”的一个主要方面是：“字”在表达概念时往往比“词”范围大。证据 1 可从“字”的“族系”现象得到证实。证据 2 可从“字”与“词（word）”的比较得到证实。证据 3 可从“字”与“word”的翻译往往需要“字组”作为中介而得到证实。证据 4 可从“字与字组（二字及二字以上的词）”之间“内涵与外延的反变关系” 而得到证实。这个特征或事实，仅从汉语词汇语义的角度看，也只有一个不是例外的例外，即“一字词”的情况。“字”的“范畴”特性，是汉语及汉语思维的一个共同特点。

确立“物、意、文、义” 四个位于“语义三棱”顶点的范畴，作为：新的“概念分类体系”的基础——总体框架，体现了汉语思维的特点。基于“语义三棱”，一方面，可从“符号分类”、“对象分类”、“专概念分类（基于：学问领域的划分，见：参考文献）”的角度，建立“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”）协同智能计算知识数据库；另一方面，可从“八大形式”与“八大关系”的角度，建立两个信息检索（简称：“两检”）系统。其中，“八大形式”系统的两个子集，即：“字”和“音”（语音部分）信息检索系统是“三注”系统数据库的两套查询系统。“八大关系”信息检索系统也是“三注”系统数据库的查询系统。也就是说，“三注”（统称:“B 库”）和“两检”（基于“A 库和 B 库”的查询系统），都是建立在同一套协同智能计算语言数据库（简称:“A 库”，涉及应用方面的进一步探讨见：“参考文献”的有关部分，如：GTCM 和 GSCM）的基础之上的。

参考文献（按照论著发表时间排序）

许国璋：论语言[C]外研社，1991

徐友渔、周国平、陈嘉映、尚杰：语言与哲学——当代英美与德法传统比较研究[M]三联书店，1996.

周斌武、张国梁：语言与现代逻辑[M]复旦大学出版社，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，1997

詹卫东，常宝宝，俞士汶：基于词组本位语法的语义模型[J]中文与东方语言信息处理学会学报 1998（1）

王路：世纪转折处的哲学巨匠：弗雷格[M]社会科学文献出版社，1998

林杏光：词汇语言学和计算语言学[M 语文出版社年，1999

Mike Beaney:1999《Meaning and Truth》[M].

鲁川：汉语语法的意合网络[M]商务印书馆，2001

李锡胤：关于“语义三角”之我见[J]俄语语言文学研究（第 1 期）2003

邹晓辉：一种知识信息数据处理方法及产品[J]发明专利公报 G06F163 知识产权出版社，2000，（11）

邹晓辉：义项语汇典例（SVDE）的总量控制模型[A]第五届汉语词汇语义学研讨会论论文集[C] 2004

邹晓辉：协同智能计算语言数据库的设计方法[J]科学知识表达的结构控制模型[J] 潜科学（第 32 期）2004（7）

邹晓辉：融智学问知识表达的基本框架体系[J]熵、信息和复杂性（第 86 期）2004，（9）

邹晓辉：融智学专著及其知识要点和基本术语（一）[J] 字的形式化定义[J]潜科学（第 38 期）2004，（12）

邹晓辉：融智学专著及其知识要点和基本术语（二）[J] 字与字组的关系[J]潜科学（第 39 期）2005，（1）

邹晓辉：解析“字与字组的关系”探索“汉语形式化”新路[J]（GTCM 和 GSCM）潜科学（第 41 期）2005，（3）

WordNet-online version；ILD-online version；Longman Lexicon of Contemporary English-online version

——从“一词泛读”到“释义字组”再到“一字精读”

摘要：本文论及一种优化“语义信息处理”的新方法。该方法源于“语义三棱”、A 库、B 库、组字公式和字组阵列。其实例是借助“释义字组”进行“一字精读”。

关键词：一词泛读、释义字组、一字精读、语义信息处理

一、绪言

本文是对汉语词汇语义学新方法的探讨，涉及：语义表示、义项的限定。特点：以“字”的“义项解析”和“释义字组”的“结构解析”作为汉语词汇语义研究的突破口。重点：精确解析“字与二字组的关系”探寻“义项”的发散与收敛的规律。研究途径：首先，根据“语义三棱”原理，把“语义信息处理”转化为“形式信息处理”。接着，对“中文形式信息”的“A 库”实施“三化”改造，再经“三注”成“B 库”。最后，通过解析“字与字组的关系”提炼出“组字公式”和“字组方阵”。本文对具体研究对象的限定：这里报道的“一字精读”示例，仅限于：对“字与二字组的关系”进行的基础性科学探讨；对“义”与“字”这两个范畴的“义项解析”示例，前者，限于“语义信息处理”及“语义三棱”的领域，后者，仅限于：本文列举的语言学主要学科分支领域；“组字公式”的示例，仅限于：由“实字”组成的“释义二字辞”。基本假设：能搞清楚“字与二字组的关系”，就能搞清楚“字与字组的关系”，进而也就能搞清楚“汉语词汇语义研究的突破口”究竟在哪里。知识贡献：指出了“字间信息”处理的基本原理，即：“组字公式”和“字组阵列”，为进一步精确地系统地解析“字与字组的关系”提供了“语义信息处理”的新方法。

二、综述

自从弗雷格开辟语言哲学的新方向，使意义问题成为哲学研究的中心问题之后，围绕着：意义的承载单位究竟是词、句子、…，还是整个语言系统，产生了不同的意义理论。

上述现象和理论，在汉语学界表现为各种“本位”说，如：“字本位”——区别于：英语的“词与句‘（双）本位’”[徐通锵（1991）]、“词组本位”[朱德熙（1982，1984，1985）]、“小句中枢”[邢福义（1995）]、“句本位”[黄昌宁（1994）]、“复本位”——区别于：“单本位”，如：“词本位、词组本位、句本位”[马庆株（1998）]。

“字与字组的关系”的研究认为：从整体上看，这些“本位”说恰似“盲人摸象”，各自仅仅摸到了（汉语这个）“大象”的一个部分。尽管如此，这仍是非常了不起的！因为（汉语这个）“大象”的确太大，致使任何个人的经历或阅历要想统观全局且一览无余都是难以想象的。

这样，狭义融智学作为一门研究自然人与计算机之间如何实现高度协作且优势互补的学问，在“人类智能”与“人工智能”之后提出的“协同智能”及其“融智概念体系”（在汉语词汇语义层面可视为：“语义三棱”模型）的功用和特长，也就必然会有用武之地。融智教学法“一字之师”的理论与实践，采用的就是“一字精读”与“字组泛读”相结合的策略。其中，“泛读”的“字组”实际上是“释义字组”。仅仅就语汇层次而言，在协同智能计算语言数据库（作为全球语言定位系统 GLPS 的实施例，简称：“A 库”）与协同智能计算知识数据库（作为全球知识定位系统 GKPS 的实施例，简称：“B 库”）中，“释义字组”有“直接呈现”与“间接呈现”两种基本形式。本研究，做了以下双向思考：从“一字精读”到“释义字组”再到“一词泛读”（由内向外与别人的研究结合）的发展；从“一词泛读”到“释义字组”再到“一字精读”（由外向内与自己的研究结合）的回顾。结果发现：1、“释义字组”用于字典词典的分析，可与“释义元语言”[苏新春（厦门大学）《论汉语释义元语言的特征》]产生“交集”。2、“释义字组”扩展到“释义句子”与“释义段落”乃至“释义篇章”，可与“一词泛读”[郑锦全（中央研究院语言学研究所）《词语管窥与宏图》]产生“交集”。探讨：如果双方都能借鉴吸收对方的研究成果，那么，可发展出什么样的互动前景呢？在计算机辅助汉语学习方面，“一字精读”或“字组泛读”

与“一词泛读”之间，可否产生互动呢？在汉语字典与词典的分析方面，“释义字组”与“释义元语言”的有关研究之间，可否产生互动呢？以下具体介绍和论述。

三、方法

优化“语义信息处理”的新方法，由四个步骤组成，即：1、应用“语义三棱”原理，实施宏观收敛；2、遵循“同义并列”法则，实施形式收敛；3、按照“三级标注”方式，实施内容收敛；4、解析“字与字组的关系”，实施微观收敛；其特征在于：

1、应用“语义三棱”原理，把“语义信息处理”转化为“形式信息处理”。

解决“语义信息处理”的问题，途径有二：一是从“概念分类体系”入手，二是从“范畴分类体系”入手。由于前者过于庞杂而后者简明扼要，所以，本研究采用后者。“范畴分类体系”是由“物（对象）、意（概念）、文（符号）、义（关系）”四大范畴构成的语义分类体系。据此，有两种“语义信息处理”策略及方法：直接处理：改进现有“概念分类体系”——基于“物（对象系统）、文（符号系统）、义（关系系统）”而构成新的优化的“概念分类体系”，其特点是：依附于人脑的“概念系统”——表现为：“思想观念”。间接处理：改进现有“语汇分类系统”——基于“物（对象系统）、意（概念系统）、义（关系系统）”而构成新的优化的“词汇分类体系”，其特点是：依附于电脑的“符号系统”——表现为：“标准文本”。

2、遵循“同义并列”法则，对“中文形式信息”的“A 库”实施“字组化、数字化、表格化”改造。

“字组化、数字化、表格化”简称：三化。构建汉语的“三化”词汇数据库，即：从“协同智能计算语言数据库”（简称：“A 库”，作为全球语言定位系统 GLPS 的实施例）中，抽取“文本总量控制模型”（GTCM）和“音节总量控制模型”（GSCM）总量相等、形式迭交的“词汇集合”，实施“字组化、数字化”改造。

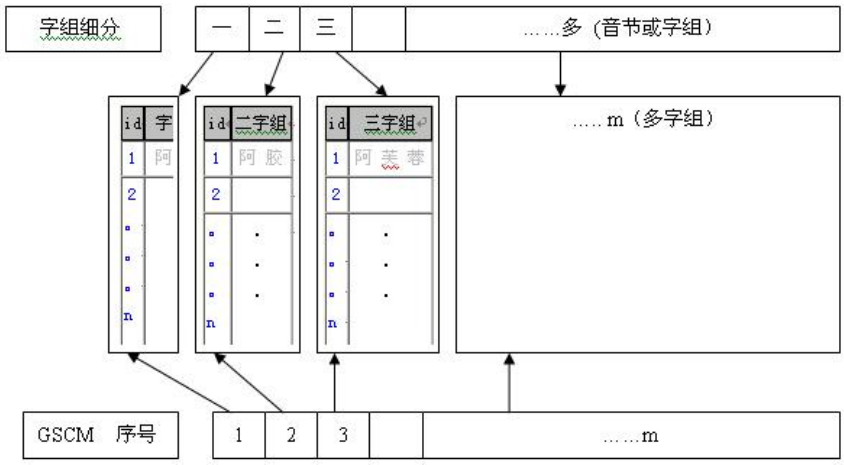


图 1

图 1 是“A 库”的“三化”示意图。

经过“三化”的“A 库”——涉及大量的“释义字组”，旨在：处理“字间信息”。其依据是通过“实字”和“虚字”在“字组”中的“序位”调节“字间的各种组合关系”的汉语语法规则和方法。

3、按照“三级标注”方式，分“三级”扩展“A 库”的“信息标注”列，构成“B 库”。

“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”。在汉语的“三化”词汇数据库的基础之上，构建“三注”知识信息数据库，即：优化的“协同智能计算知识数据库”（简称：“B 库”，作为全球知识定位系统 GKPS 的实施例）。

图 2

图 2 是“B 库”的“三注”示意图。

经过“三注”的“B 库”——同时可提供“一词泛读”的语料，旨在：处理“字外信息”。其依据一方面是“三注”信息，如：通用常识与科学知识两个系列的分类信息标注；另一方面是语言文字的分类信息标注，如：基于“语义三棱”原理的“基本语义语法分类”，如：实字及实字组部分的“虚的对象的称谓与实的对象的指称”、“抽象的属性与直观的状态”、“瞬间的动作与连续的过程”、“静的机理与动的法则”；虚字及虚字组部分的“近距关系与远距关系”。

4、解析“字与字组的关系”。

4.1、把“释义字组”限定在“二字组”的范围，4.2、用“字”的“义项解析”和“释义字组”的“结构解析”相结合的方式，做“一字精读”。4.3、重点研究“字与二字组的关系”。

四、结果

以下从理论结果、实践结果和应用结果三方面介绍：新方法的功用。

1、新方法在理论上的收敛步骤及预期结果

1.1、宏观收敛的结果：由“语义三棱”的四个顶点向一个顶点聚焦，即：由“物（对象）、意（概念）、义（关系）”三个范畴向“文（符号）”一个范畴进行收敛。

1.2、形式收敛的结果：由“字内形式信息处理、字间形式信息处理、字外形式信息处理”乃至“字里行间的形式信息处理”向“字间形式信息处理”进行收敛。

1.3、内容收敛的结果：由“字内内容信息处理、字间内容信息处理、字外内容信息处理”乃至“字里行间的内容信息处理”向“字间内容信息处理”进行收敛。

1.4、微观收敛的结果：由“多字间信息处理”向“二字间信息处理”进行收敛。

2、新方法在实践上的收敛步骤及实际效果

众所周知，“字”集中反映了汉语的特点。“什么是字？”汉语理论界至今无定论。我们认为：这样的提问方式不是一个便于操作的方式。因为，单单一个“字”，其含义，既模糊，又不确定，各种可能性都存在。“什么是字？”或“字是什么？”的问题，恰似“语义三棱”理想实验“设想 1”的情形，即：“四点合为一点”的情形。结果只能是：什么都是，什么又都不是。

为此，不如换一种便于操作的提问方式。如：汉语“字本位”理论所说的“字”究竟涵盖了什么重要的语言学信息？要完整地回答这个问题，虽然也不是一件容易的事，但是，毕竟具有可操作性。

下面尝试应用本文所述的理论和方法对这个具有挑战性的汉语“字本位”理论的一个根本问题给出自己的答案或收敛步骤及结果。我们就以“语义三棱”原理“实验 2”的方式或途径，作为正式回答这个问题的“切入点”。以“字”为例来检验上述新方法的收敛效果。

2.1、宏观收敛的效果：“字”属于“文（符号）”这个范畴，位于“语义三棱”四个顶点中的一个。“文（符号）”范围还是太宽，其中，“字内信息”通常属于文字学的研究范围，“字间信息”既涉及语义学又涉及语法学的研究范围，“字外信息”通常属于语用学的研究范围，“字里行间的信息”既涉及词汇学又涉及语音学的研究范围。对“汉语”而言，以“字”作为“基本

语言结构”具有独特的功效，既涉及“字内”又涉及“字间”甚至还涉及“字外”（如：话外音）的信息处理。

2.2、形式收敛的效果：“字”可能涉及“字间形式信息处理”。如：从“A 库”中找出“字”所在的“表”。

2.3、内容收敛的效果：“字”可能涉及“字间内容信息处理”。如：从“B 库”中查出“领域信息”。

2.4、微观收敛的效果：由“多字间信息处理”向“二字间信息处理”进行收敛。

从“A 库”中“二字表”中查出：含有“字”的“二字组”。根据“B 库”中“字”的“领域信息”选出“可以作为语言学主要分支学科的微观研究对象”的“字”的“字间与字外信息”。具体步骤和结果如下：

2.4.1、从“A 库”中查出“字”的“前字”和“后字”信息——“字间信息”。

2.4.1.1、“字”，作为：被其他字限定其义项范围的“解字”。如：铸字、正字、正字、脏字、许字、虚字、习字、文字、题字、题字、俗字、数字、熟字、实字、识字、生字、如字、签字、铅字、排字、名字、盲字、活字、画字、汉字、方字、点字、单字、待字、打字、错字、赤字、衬字、拆字、测字、草字、别字、表字、本字、白字、八字、……

2.4.1.2、“字”，作为：去限定其他字的义项范围的“用字”。如：字帖、字谜、字面、字眼、字幕、字母、字体、字书、字样、字模、字典、字汇、字号、字画、字据、字迹、字调、字句、字纸、……

可见，即使“二字组”可与其搭配的范围也相当宽。接着需对字的义项和释义字组实施“领域”限制。

语言学分支	研究对象（微观部分）	
汉语文字学	字形	
汉语语音学	字音	语音，语调
汉语语汇学	字汇，辞汇，语汇	
汉语字典学	字	
汉语辞典学	字组，辞，语	（语词）
汉语语义学	实字 字义	
汉语语法学	虚字	语序（标点符号）
汉语语用学	前字 后字	语句 （上下文）

图 3

图 3 是语言学主要分支学科的微观研究对象排列。

2.4.2、从“B 库”中查出“领域信息”——“字外信息”。

2.4.3、解读“可作为语言学主要分支学科的微观研究对象”的“字”。

3、解析“字与字组的关系”的过程中提炼出“组字公式”与“字组阵列”。汉语“字本位”理论，把语汇分为“字、辞、块”三种基本类型。这里把“辞”与“块”统称为“语”。这样，“组字成语”的逆过程就是“分语为字”，其中，涉及：切“辞”、分“块”（即：从“语”中切分出“辞”与“块”）两个步骤。在字与（各级）字组的关系中，字与二字组的关系是基础，下面给出“二字组”基本关系的科学描述。如果把需解释或限定其义项的“字”命名为“解字”，把限定“解字”义项的“字”命名为“用字”，那么，限于“二字组”的“释义字组”就只有“释辞”与“释语”两种类型。

“释义二字组”=“用字”+“解字”。这是：基本组字公式。

3.1.“释辞”=“实字”+“实字”=“用字”+“解字”。这是“基本释辞公式”。

3.2.“释语”=“虚字”+“实字”=“用字”+“解字”；

3.3.“释语”=“实字”+“虚字”=“用字”+“解字”；

3.4.“释语”=“虚字”+“虚字”=“用字”+“解字”。

上述 3.1.-3.4.四个公式中的实字与虚字的关系，恰似一个阵列。故简称：字组阵列，也是：基本字组阵列。后续的“三字组”、“四字组”、…、“多字组”的“字组阵列”都可基于这个“基本字组方阵”的原理而推衍出来。同理，后续的“三字组”、“四字组”、…、“多字组”的“释辞公式”也都可基于这个“基本释辞公式”的原理而推衍出来。同理，后续所有的“组字公式”都可基于这个“基本组字公式”的原理而推衍出来。例如：“释义三字组”也有“释辞”与“释语”两种类型。其中，

“释辞”=“实字”+“实辞”=“用字”+“解辞”；

“释辞”=“实辞”+“实字”=“用辞”+“解字”；

“释语”= … …。进一步的划分或枚举，由于可类推（如：“用语”、“解语”）和可枚举（如：“四字组”、“五字组”的“释义字组”），所以，本文不再做具体介绍。

4、新方法的直接应用结果

无论是从“一字精读”到“一词泛读”，还是从“一词泛读”到“一字精读”，其中，都离不开“字组”或“释义字组”这个“中介”。下面就从“字组”的解析，看“一字精读”与“一词泛读”的互动关系。

4.1、应用实例：“义”与“语义”的探讨

“语义信息处理”这个“六字组”，既可由“语义信息”与“信息处理”二个“四字组”交错排列而成，也可由“语义”、“信息”和“处理”三个“二字组”顺序排列而成。前三个“字组”是“词组”，后三个“字组”是“词”。习惯上，分析也就到此为止了。什么是“语义”？或“语义”涵盖多少信息？仍不清楚。

根据“语义三棱”——“物、意、文、义”的范畴划分法，“语义”一词还可分解为“语”与“义”这两个各自涉及一组“义项”的“字”——更基本的“语言结构单位”。其中，“语”字涉及的一群“概念”，可由两个系列的“二字组”展示如下：“语词、语句、语段、语篇、…”和“话语、母语、外语、…”。“义”字涉及的一群“概念”，可由后面的“二字组”系列呈现如下：“**义理**、…”和“**本义**、同义、近义、反义、褒义、贬义、广义、狭义、含义、歧义、辨义、定义、音义、意义、法义、…”。

4.2、探讨结果：发现或证实“字”与“词（word）” 在内涵上的差异

上述对“语义”这个“二字组”的解析，得到一个结果：由汉语的“（本土）字”融合“（外来）词”而重构的“（本土）辞”显然比其对应的“（外来）词”本身增加了更丰富的内涵。由此，发现或证实：汉语的一个非常重要的特点，即：“字”的义项“组群化”或概念“范畴化”。

4.3、“双语存储”模型和“双语语义”词典，对“字词（word）差异”的词汇处理

根据认知心理学“双语存储”模型——“共同存储说”与“单独存储说”，我们提出了涵盖前两个模型（作为特列处理）的“协同存储说”。基于“A 库与 B 库”，我们构建了：基于“协同存储”**模型**的**“双语语义”词典**。可直接用于解释和构建：汉语的“（本土）字”融合“（外来）词”而重构“（本土）辞”的认知存储模式和**“双语数字化”系统**。基于“语义三棱”我们对认知的“对象、概念、符号、关系”究竟是“共同”拥有还是“单独”具有，把“双语对译”的“认知存储模式”分为三种基本类型：**4. 3.1、**基于“共同存储说”的“同义并列”类型，如：天（sky）、水（water）、盐（salt）、…（直接对译）。**4.3.2、**基于“单独存储说”的“借鉴重组”类型，如：电报、电话、计算机、**信息处理**、…（汉语直接吸收英语传递的信息）；china（瓷器）、panda（熊猫）、…（英语直接吸收汉语传递的信息）。**4. 3.3、**基于“协同存储说”的“借鉴重组”类型，如：**语义信息**、油画、水彩画、…（交互融合迭交共现）。基于上述原理和方法开发的字典词典系统，在字汇词汇的实字实词范围以内，可相应地建立汉语与英语的“对译关系”：部分“（本土）字”与部分“（外来）词”之间的“对译关系”，部分“（本土）辞”

与部分“（外来）词”或部分“（外来）词组”之间的“对译关系”。

五、结论

结论 1：新方法及其系列工具（如：A 库与 B 库）具有系统精确的收敛功用。

“四个步骤”依次递进的过程，同时，也是：逐步收敛的过程。对“双语语义”词典以及“双语数字化”系统而言，收敛精准=消歧=无争议。一个“字”有多少信息？四步收敛之后，问题也就清楚了。

结论 2：“组字公式”和“字组方阵”，可作为解析“字与字组的关系”的微观分析工具。

说话的过程，就是在根据谈话的需要，不断地限定“字”的“义项”的过程。这个“限定过程”由“一连串的字”组成。字与字的关系，由“语序”和“虚字”来调节。凡是“实字”与“实字”能直接“接续”的关系，就由“语序”调节；凡是不能直接“接续”但能间接“联系”的关系，就加入“虚字”来参与调节。直接“接续”与间接“联系”是字间关系的两种基本形式。“组字公式”和“字组方阵”对此极有用。

结论 3：汉语“语义信息处理”以处理“字内信息”和“字间信息”为基础。

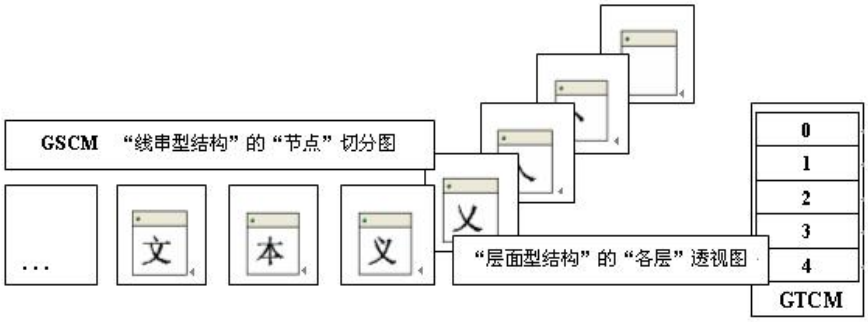


图 4

图 4 是：“字内信息”与“字间信息”示意图。

图 4 既揭示了“字内信息”处理机制，又揭示了“字内信息”与“字间信息”的关系。“字”是“线串型结构”与“层面型结构”的“迭交”形式。“字间关系”就是“线串型结构”的各个“节点”（含：“起点”）之间的关系。“字内关系”则是“层面型结构”的各个“层面”（以透视的方式由“顶层”可见“各层叠合”的字形）之间的关系。“字内信息”表现“层面型结构”的“字内关系”；“字间信息”表现“线串型结构”的“字间关系”。汉语的语义信息，主要是“字内信息”与“字间信息”（“字外信息”通常视为:语用信息）。汉语的语法信息，通常只是“字间信息”。由此可见，“字间信息”是语义信息与语法信息之间的交集。“字间关系”有直接与间接之分，也有字组与字串（带标点符号的句子）之分，…。记录语义的形式，可有：记录词汇语义的字组与记录句子语义的字串。汉语的语法，从根本上说，就是“语序”和“虚字”调节“字间关系”的方法或规则，其基础就是：“字与二字组的关系”。“组字公式”与“字组阵列”有利于：清理各种“语序”和归纳“虚字”用法（其中表现了：对“字义”或“字的用法”的各种“限定方式”）。

总结：“语义三棱、A 库、B 库、组字公式和字组阵列”结合应用，可优化“语义信息处理”。

六、议论

几个有待进一步探讨的问题：

1、“组字公式”的两组用词，那一组更好？

“组字公式”=“用字”+“解字”=“前字”+“后字”。例 1：“用字”+“解字”=“释辞”。例 2：“前字”+“后字”=“二字组”。例 1 与例 2 虽可表达同样的公式但精确程度不同。“用字、解字、释辞”是专用术语。能突出汉语与英语的区别。不习惯的读者可暂时采用旧词或旧概念，虽不精确，但也有近似效果。

2、什么是汉语“字本位”理论的“字”？我们给出的答案是：

答案 1：“音、形、义”三合一的“字”，是组成“字组”以及“字串”等“线串型结构”的“基本结构单位”。从“语义信息处理”的角度看，以“字”为界线，划分：“字内信息”、“字间信息”和“字外信息”便捷高效。就“字间信息”处理而言，以“字”为单位，划分：一、二、三、…、多“字”的“线串型结构”对“三化”词汇数据库的构建和应用都十分方便。“组字公式”和“字组阵列”的系统化应用，既有利于“汉语形式化”，又有利于系统地整理“汉语语法”体系。

答案 2：“音、形、义、用”四合一的“字”，或“音、形、义、解”四合一的“字”，是对“答案 1”进一步发展。其中，涉及语言学主要分支学科的一系列微观研究对象——“音字、形字，实字、虚字，用字、解字”的探讨。尤其是“音字、形字，用字、解字”等概念，有待进一步探讨。

答案 3：在“答案 1”或“答案 2”基础上探讨“字”的特点：字的形式，在“线串型结构”中处于“起点”或“节点”的地位，呈“迭交”状态，即：“线串型结构”与“层面型结构”的“迭交”；字的内容，即：义项，在语义上（与字组相比）表现出：义项“组群化”或概念“范畴化”的倾向。在语法上表现为：实字、虚字，用字、解字，…。字的义项，与“前字”结合，呈：收敛性；与“后字”结合，呈：发散性。

一个附带的问题：就形态而论“音与形”迭交的“音字”（如：“形声字”的“声符”本身就是 1325 个“音字”）比“拼音与字形”分离的“形字”更具独特性。以上就是我们对“字”的研究结论。

3、语义语法的基本分类，如：实字及实字组部分的“虚的对象的称谓与实的对象的指称”、“抽象的属性与直观的状态”、“瞬间的动作与连续的过程”、“静的机理与动的法则”；虚字及虚字组部分的“近距关系与远距关系”，…。

4、“字”与“词”的异同：通过“一字精读”与“一词泛读”，比较“字”与“词”的异同。可得出“汉语形式化”体系设计和“计算机辅助”词汇教学“精泛并举、优势互补”的结论。通过“字间信息”处理，探讨“字词关系”中进一步必然涉及的两个重要关系，即：一、“字与范畴”和“词与概念”的关系；二、“字与组字”与“字与词”的关系。与“词（word）”比较而言，“字的义项发散性与范畴化特征”及“字组的义项收敛性与概念化特征”是汉语以及汉语思维的一个基本特点。

希望本文能给读者一些有益的参考或启示。欢迎学界同行提出宝贵意见！

参考文献（按照论著发表时间排序）

陈肇雄主编：机器翻译研究进展[C] 1-564 页，电子工业出版社，1992

方　立：美国理论语言学研究[M]1-240 页，北京语言学院出版社，1993

石　锋：汉语研究在海外[M]123-188 页，北京语言学院出版社，1995

汪安圣等：认知心理学[M] 344-367 页，北京大学出版社 1996

徐通锵：语言论--语义型语言的结构原理和研究方法[M] 295-442 页，东北师范大学出版社， 1997

陆俭明、郭锐：汉语语法研究面临的挑战[J 世界汉语教学，1998，（4）

詹卫东，常宝宝，俞士汶：基于词组本位语法的语义模型[J]中文与东方语言信息处理学会学报 1998（1）

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] 1-516 页，清华大学出版社，1998

林杏光：词汇语言学和计算语言学[M60-118，140-376，语文出版社年，1999

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

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

鲁　川：汉语语法的意合网络[M]1-277 页，商务印书馆，2001

邹晓辉：一种知识信息数据处理方法及产品[J]发明专利公报 G06F163 知识产权出版社，2000，（11）

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]41-44 页，商务印书馆 2002

俞士汶：关于汉语信息处理的认识及其研究方略[J]语言文字应用（总第 42 期）2002，（2）
王开杨：“一语双文”的理论基础和面临的困难[A]见苏培成等编：语文现代化论文集[C] 商务印书馆，2002
王　惠：汉英机器翻译中基于大型语义词典的汉语词义消歧[A]黄河燕主编：机器翻译研究进展[C]电子工业出版社，2002
邹晓辉：语言及语义信息的统一参照系[J]潜科学[诸子百家：邹晓辉的融智学（系列文章之一）]2002.05
郑锦全：词语管窥与宏图[A]第五届汉语词汇语义学研讨会论论文集[C] 2004
苏新春：论汉语释义元语言的特征[A]第五届汉语词汇语义学研讨会论论文集[C] 2004
邹晓辉：义项语汇典例（SVDE）的总量控制模型[A]第五届汉语词汇语义学研讨会论论文集[C] 2004
邹晓辉：协同智能计算语言数据库的设计方法[J]潜科学（第 32 期）2004（7）
邹晓辉：论汉语字组的细分[J]潜科学（第 32 期）2004（7）
邹晓辉：汉语“字本位”理论研讨会论论文：字的形式化定义[J]潜科学（第 38 期）2004，（12）
邹晓辉：汉语“字本位”理论研讨会论论文：字组划分数字化[J]潜科学（第 38 期）2004，（12）
邹晓辉：汉语“字本位”理论研讨会论论文：字与字组的关系[J]潜科学（第 39 期）2005，（1）
邹晓辉：协同智能计算知识数据库的设计方法[J]潜科学（第 39 期）2005（1）
邹晓辉：解析“字与字组的关系”探索“汉语形式化”新路[J]潜科学（第 41 期）2005，（3）

全国第八届计算语言学联合学术会议(JSCL-2005)交流论文

中文信息处理的新方法

摘要：为实现自然人与计算机的优势互补，通过人机分工协作，我们提供了中文信息处理的新方法。主要涉及“文本总量控制模型”、“中国标准信息交换码”和“音节总量控制模型”。

关键词：语言数据库、知识数据库、间接形式化、中文信息处理

绪言

本文涉及：知识表示、语料库语言学、记忆模型、机器学习、知识获取和推理技术等，属于计算语言学基础研究领域。其特点是：采用“间接形式化”方法，直接进行“中文信息处理”。其重点是：通过自然人与计算机之间的分工协作而实现优势互补。研究途径：借助“关系数据库”的形式（涉及：前台界面、后台程序以及前后台语言转换）实现“（整型）数字”与“（字符串）文字”的“同义并列”达到“人机协作且优势互补”的目的。其局限是：人机协作必须依据共同的“参照系”。基本假设：基于“关系数据库”而实现的“数字”与“文字”的“同义并列”为人机协作提供的数理“参照系”，是实现由非线性的中文信息处理转化为线性的数字信息处理的依据。知识贡献：不仅给自然语言处理找到了一条基于母语的“间接形式化”方法——如：使非线性的中文信息处理转化为线性的数字信息处理，而且，为“人类智能”与“人工智能”之后提出的“协同智能”提供了一个典型实例——即：汉语的“间接形式化”示例（既区别于又兼容于通行的“形式化”——基于英语的“间接又间接的形式化”）。希望本文介绍的“中文信息处理的新方法”能给学界同仁有所助益或启示！

综述

众所周知，通行的汉语“形式化”探索一直以来没有突破这样一种格局：

一方面，汉语至今没有自己独立的“形式化”方法，而不得不建立在基于英语的“形式化”基础之上。这就必然带来以下几个问题：1、两种“自然语言”之间的“不可翻译”部分，必然造成一系列隐含的问题。2、两种“思维方式”之间的“根本区别”部分，难以通过“翻译”消解分歧。习惯“汉语思维”的人，即使英语掌握得较好，也会面临：“双语”冲突问题。3、“英语处理”仍存在“形式化”问题。目前，所谓基于英语的“形式化”，实质上是基于英语字母的各种人工语言的“形式化”。“英文信息处理”仍需“程序语言”为“中介”才能进行。4、目前，通用计算机中借助“汉化”、“翻译”或“解释”而进行的所谓汉语的“形式化”探索，从来就没

有离开过上述 1、2、3 条限制，其困难和受制约程度是不言而喻的。

另一方面，汉语基础理论，长期没有一个可有效解释各种语言事实的完整理论或总体框架，特别是：汉语语法的问题长期争论不休而得不到满意的解决方案。这又进一步增加了以下问题：5、作为自然人的汉语专家都分辨不清的“语法语义分歧”又如何让计算机重用。6、即使作为自然人的汉语专家已分清的“语法语义分歧”至今也未全部实现计算机重用。

事实证明：无论是基于现代逻辑的“符号集”，还是基于“美国标准信息交换码”的“字符集”，甚至是在其基础上形成的各种“计算机程序语言”，都与“汉语体系”、“汉语思维方式”和“汉语字符集”无直接渊源关系。现在学界进行的所谓汉语的“形式化”探索，实际上，不仅隔着英语，而且还隔着一系列“中介语言”（如：基于英语的各种程序语言——人工语言）及其“符号法则”（如：基于程序语言的形式语法）。真可谓：间接又间接的“间接又间接的形式化”。其困难程度由此也可见一斑。

从发现这个根本问题至今，笔者做了一系列的尝试。先后提出了以下设计方案：一种智能通信子母机，一种知识信息数据处理方法及产品，协同智能计算语言数据库的设计方法，协同智能计算知识数据库的设计方法，义项语汇典例（SVDE）的总量控制模型，优化“语义信息处理”的新方法与实施例，解析“字与字组的关系”探索“汉语形式化”新路。实践证明：基于母语的“间接形式化”方法是切实可行的。现以“中文信息处理”为例“汇总、提炼、概括”如下：

方法

本文介绍的中文信息处理新方法，涉及以下几个基本步骤：

1、划分“子全域”、“超子域”和“进阶层式”；

2、借助“关系数据库”构建“总量控制模型”；

3、区分“文本”和“音节”两类“汉语模型”；

4、借助该模型实现汉语的“三化”及“三注”，

其特征在于：

1、“进阶层式”各表的“文字”与“数字”是“同义并列、一一对应的关系”；

2、“进阶层式”的各个“一览表”都有各自“唯一的表号 m_i 作为：序列代码”；

3、“进阶层式”的各个“一览表”中的“文字与数字”的“序位均同义并列”；

4、无论“文本”还是“音节”的汉语模型“序位代码（m_i,n_j）都是唯一的”；

5、（m_i,n_j）构成的“矩阵”涉及“线性方程组”和各种各样已知的“好算法”；

6、基于上述数学模型及关系数据库容易实现基于母语的程序设计和自动重用；

7、体现“字与字组的关系”的“组字公式和字组阵列”便于“人机协作消歧”。

上述方法涉及的概念或术语，见：**注释**。

结果

上述中文信息处理新方法，是直接基于算术“数字”和汉语“文字”之间显性分工协作的“（汉语）间接形式化”方法，不仅区别而且兼容于：直接基于算术“数字”和英语“字母”的各种程序语言之间隐性分工协作的“（英语）间接又间接的形式化”方法，同时，也不仅区别而且兼容于：（汉语）间接又间接的“间接又间接的形式化”方法。

实施中文信息处理新方法，则产生以下结果：

1、语言总量控制模型——“A 库”（本文仅仅介绍其汉语部分）

“GTCM”（“文本总量控制模型”，见：图 1）是“超子域”及其各“进阶层式”的汉语“间接形式化”体系，其中，涉及两个特殊的部分，一个是“子全域”——“Z-ASCII”（“中国标准信息交换码”）（见：图 2）；另一个是词汇一级“粗分子模型”的并列“细分子模型”——“GSCM”（“音节总量控制模型”，见：图 3）。对“中文信息处理”而言，根据“字内信息”、“字间信息”或“字外信息”，“GTCM”可分为以下三组“一览表”，即“m_i”分别取值为：0-4，4-6，5-12。由“GTCM”的“4-6” 三个“一览表”合并之后，再按“字本位”重组成为（GSCM）“1- m”个“一览表”，

着重分析“字间信息”。在此，“GTCM”是广义的“汉语形式化”模型；“GSCM”是狭义的“汉语形式化”模型。

2、知识总量控制模型——“B 库”（本文仅仅涉及其概念部分）

基于“A 库”的“B 库”是按“三注”的方式，分三级扩展为多组信息标注“列”而构成的“知识信息查询”数据库。基于标准化的“A 库”和“B 库”(见：图 4)，用户可建构个性化“N 库”(“软件总量控制模型”)。

图 1 是文本总量控制模型(涉及 GTCM 的 0-12 个表)示意图。图 2 是中国标准信息交换码(Z-ASCII 仅涉及 GTCM 的 0-4 个表中的第一个表）示意图。图 3 是音节总量控制模型（涉及 GSCM 的 1- m 个表）示意图。图 4 是基于“A 库”的 GSCM 经“三注”后成为“B 库”示意图。见：**附图**。

结论

1、Z-ASCII 可带来计算机底层技术的原创性实质突破

通过 GTCM 可构建基于汉语且兼容英语的“中国标准信息交换码”——Z-ASCII。如：把 Unicode 中汉字的“单面型固定结构”改为“层面型活动结构”，实现“子全域”与“超子域”及其各“进阶层式”一览表之间清晰的形式划分，完成汉字的“模糊信息处理”向“清晰信息处理”的转化。

2、GTCM 及 GSCM 可带来汉语（文字、语音、语义、语法、语用）信息处理的原创性实质突破

以汉语的“GTCM”即“超子域”第 4“进阶层式”单音节的“字”为“基本结构单位”构建后续第 5-8“进阶层式”多音节的“字组”。其中，第 4-6 段，涉及：GSCM，属于：词汇及词法一级的“信息处理”；第 7-8 段，属于：句子及句法一级的“信息处理”。也就是说，第 4-8 段的词语搭配的“约束条件”的“形式集合”就构成了“汉语（文字、语音、语义、语法、语用）信息系统”。

3、GTCM 及 GSCM 的“表格化”形式体系的优点：人机交互的“母语化”

就可使用母语编程而言，优于：各种程序语言。如：普通用户也可直接使用汉语编程（“人助机”的过程），相当于“母语化”的 SQL（结构化查询语言）和 XML（可扩展标记语言）以及其它各种常用的程序语言，歧义则由“母语化”的数学语言直接排除（“机助人”的过程）。

议论

1、具体构建 Z-ASCII 的几个可选方案

a、采用四个“理想笔画”直接与 ASCII（美国标准信息交换码）兼容的方案。**b、**采用二十七个“基本笔画”间接与 ASCII 兼容的方案。**c、**“笔画”选定后，还要选择“软、硬、软硬结合”组件的具体转换方式。

2、具体构成“汉语（文字、语音、语义、语法、语用）信息处理系统”的几个可选方案

a、GTCM 的“0-4”表中分解的“字内信息”是否要利用？笔者认为应该利用，而且，GTCM 和 Z-ASCII 有条件利用。然而，现在通行的“字处理标准（如：基于 ASCII 的 GBK 和 Unicode）字处理方式（如：FONTS 字库）”没有也无法利用。**b、**GTCM 的“4-6”表中分述的“字间信息”是否“可相对完全归纳”？实验证明：可以。**c、**总量上等于 GTCM 的“4-6”表的“GSCM”的“1-m”表的“字-字组”以及“字间信息”是否“可穷举”？实验证明：在“相对完全归纳”的条件下“可穷举”——可做到“集大成”。**d、**GTCM 的“5-12”表中分述的“字外信息”是否“可穷举”？实验证明：“5-8” 表中分述的“词法句法信息”、“9-10”表中分述的“章法信息”与“11-12”表中分述的“分类编目信息”在“相对完全归纳”的条件下“可穷举”——可做到“集大成”。**e、**GTCM 的“0-12”表中分述的“中文信息”以及“GSCM”的“1-m” 表中分述的“汉语信息”是否“可穷举、可贯通”？实验证明：在“相对完全归纳”的条件下，借助“三化”和“三注”的条理化信息处理方式，“可穷举、可贯通”。现在通行的做法是分离的——各自为政、一盘散沙，如：基于 GBK 和 Unicode 及 FONTS 中的汉字的“单面型固定结构”，与“4-6”表对应的“电子词典”以及“分词与标注”或“词法分析”，与“7-8”表对应的“句法分析”，与“9-10”表对应的“章法分析”，与“11-12”表对应的“分类编目”——即使采用现行的“数字图书馆”方

案，也因为没有涉及解决本文所述的“汉语形式化”的基础性问题而难以“条理化地穷举、贯通”。

综上所述，组织巨型的“汉语信息处理”或“中文信息处理”系统工程，有一系列具体工作要做。而目前最重要的应是“间接形式化”与“间接又间接的形式化”两种标准或道路的选择问题。

注释：

1、“子全域”，指：基于汉语且兼容英语的“中国标准信息交换码”（Z-ASCII），是“协同智能计算语言数据库”即“文本总量控制模型（GTCM）”第一个基础表的“基准符号集”。2、“超子域”，指：基于 Z-ASCII 的后续“组合符号集”，涉及“协同智能计算语言数据库”即“文本总量控制模型（GTCM）”第二至十二个基础表，其中，第五至七个基础表经“字本位”重组可对应地转换为“音节总量控制模型（GSCM）”第 1 至 m 个基础表。3、“进阶层式”，对汉语而言，有两组，即：GTCM 第零至十二共 13 个基础表，其中，第五至七个基础表等价于 GSCM 第 1 至 m 个基础表。4、自然语言处理的“总量控制模型”，对汉语而言，就是：GTCM 第零至十二共 13 个基础表与 GSCM 第 1 至 m 个基础表。5、汉语的“文本总量控制模型”，即：GTCM 第零至十二共 13 个基础表。6、汉语的“音节总量控制模型”，即：GSCM 第 1 至 m 个基础表。7、“三化”，字的定义表格化，字组划分数字化，义项呈现字组化。8、“三注”，语言文字信息标注，通用常识信息标注，专用知识信息标注。9、“双语”，有广义与狭义之分。狭义的“双语”，如：以“字母”为“子全域”的“英语”与以“笔画”为“子全域”的“汉语”。广义的“双语”，如：以“数字”为“子全域”的“算术语言”与以“字符”为“子全域”的“自然语言”。10、“m_i”表示：表号；“n_j”表示：格号。都取“自然数”的“值”。11、“好算法”是纯数学术语，区别于“坏算法”——导致“指数爆炸”的算法。12、“字与字组的关系”，涉及：形式与内容两方面，其中，基础是：字与二字组的关系。13、“组字公式和字组阵列”，汉语“字本位”理论，把语汇分为“字、辞、块”三种基本类型。这里把“辞”与“块”统称为“语”。这样，“组字成语”的逆过程就是“分语为字”，其中，涉及：切“辞”、分“块”两个步骤——从“语”中切分出“辞”与“块”。鉴于“字与字组的关系”中“字与二字组的关系”是基础，在此主要给出“二字组”类型的“辞”与“块”的“组字公式和字组阵列”。如果把需解释其义项的“字”命名为“解字”，把限定“解字”义项范围的“字”命名为“用字”，那么，限于“二字组”的“释义字组”就只有“释辞”与“释块”两种类型。

组字公式 1.**“释辞”**=“实字”+“实字”=“用字”+“解字”。

组字公式 2.**“释块”**=“虚字”+“实字”=“用字”+“解字”；

组字公式 3.**“释块”=**“实字”+“虚字”=“用字”+“解字”；

组字公式 4.**“释块”**=“虚字”+“虚字”=“用字”+“解字”。

上述 1.－4.**四个公式**中实字与虚字的关系，恰似一个阵列。故简称：字组阵列。后续“三字组”、“四字组”、…、“多字组”的“组字公式和字组阵列”都可基于上述“二字组”的基本“组字公式和字组阵列”的原理推衍出来，故不再做具体介绍。

14、“人机协作”，有隐性与显性两种分工协作形式，无论是“实时、分时和批处理”还是“计算机集中处理与网络分布处理”均可采用。

当前通行的那种借助“关系数据库”的形式对“（字符串）文字”的“同义并列”的“人与机之间分工协作”方式是“隐性的”，即：前台界面、后台程序以及前后台转换的“中介语言”及其“法则”是多样化的，如：各种“程序语言”及“形式文法”——因此从事汉语信息处理的专家至少必须熟悉“汉语、英语、程序语言、数学”四方面的多种知识技能，这恰似“旧式的全能裁缝”，因此，也难以组织巨型的“中文信息处理”系统工程；

本文所述的这种借助“关系数据库”的形式对“（字符串）文字”与“（整型）数字”的“同义并列”的“人与机之间分工协作”方式是“显性的”，即：前台界面、后台程序以及前后台转换的“中介语言”及其“法则”是唯一的，因此，从事汉语信息处理的专家通常只须熟悉“汉语”知识技能的一个方面，这恰似“新式的流水作业”，便于组织巨型的“中文信息处理”系统工程。

附图：

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 1

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	亠	乙	阿
3	丿	冂	冫	二	啊
4	丶	勹	口	十	锕
5	乙	冖	氵	厂	錒
6	…	…	…	…	…

图 2

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 3

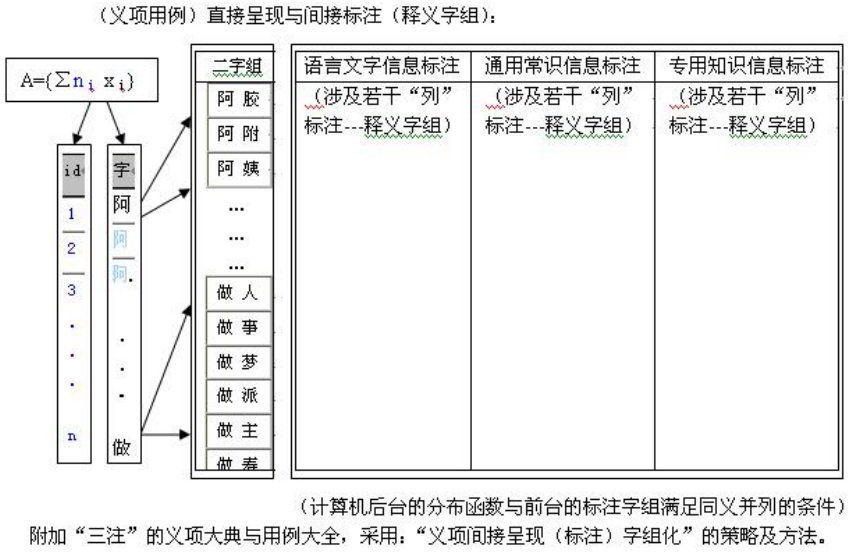


图 4

参考文献：

熊全淹：近世代数[M] 15-120 页，上海科学技术出版社，1978

中国人民大学数学教研室：线性代数[M]85-138 页，1983

陈肇雄主编：机器翻译研究进展[C] 1-564 页，电子工业出版社，1992

方　立：美国理论语言学研究[M]1-240 页，北京语言学院出版社，1993

喻云根：英汉对比语言学[M] 69-99 页，北京工业大学出版社，1994

朱志凯：逻辑与方法[M]3-32，225-287，229-304 页，人民出版社，1995

石　锋：汉语研究在海外[M]123-188 页，北京语言学院出版社，1995

王　甦、汪安圣：认知心理学[M] 344-367 页，北京大学出版社 1996

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]语言学论辑[C] 1-17 页，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] 295-442 页，东北师范大学出版社， 1997

陆俭明、郭锐：汉语语法研究面临的挑战[J 世界汉语教学，1998，（4）

詹卫东，常宝宝，俞士汶：基于词组本位语法的语义模型[J]中文与东方语言信息处理学会学报 1998（1）

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] 1-516 页，清华大学出版社，1998

林杏光：词汇语言学和计算语言学[M60-118，140-376，语文出版社年，1999

俞士汶、朱学锋：计算语言学文集[C] 1-254 页，北京大学计算语言学研究所，2000

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

康博创作室：SQL Server 2000 数据仓库设计和使用指南[M] 14-69，113-230 页，清华大学出版社 2001

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

鲁　川：汉语语法的意合网络[M]1-277 页，商务印书馆，2001

邹晓辉：一种知识信息数据处理方法及产品[J]发明专利公报 G06F163 知识产权出版社，2000，（11）

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]41-44 页，商务印书馆 2002

俞士汶：关于汉语信息处理的认识及其研究方略[J]语言文字应用（总第 42 期）2002，（2）

王开杨：“一语双文”的理论基础和面临的困难[A]见苏培成等编：语文现代化论文集[C] 商务印书馆，2002

黄河燕主编：机器翻译研究进展[C] 1-282 页，电子工业出版社，2002

邹晓辉：语言及语义信息的统一参照系[J]潜科学 2002.05

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

郑锦全：词语管窥与宏图[A]第五届汉语词汇语义学研讨会论论文集[C] 2004

苏新春：论汉语释义元语言的特征[A]第五届汉语词汇语义学研讨会论论文集[C] 2004
邹晓辉：义项语汇典例（SVDE）的总量控制模型[A]第五届(国际)汉语词汇语义学研讨会论论文集[C] 2004
邹晓辉：协同智能计算语言数据库的设计方法[J]潜科学（第 32 期）2004（7）
邹晓辉：论汉语字组的细分[J]潜科学（第 32 期）2004（7）
邹晓辉：汉语“字本位”理论研讨会论论文：字的形式化定义[J]潜科学（第 38 期）2004，（12）
邹晓辉：汉语“字本位”理论研讨会论论文：字组划分数字化[J]潜科学（第 38 期）2004，（12）
邹晓辉：汉语“字本位”理论研讨会论论文：字与字组的关系[J]潜科学（第 39 期）2005，（1）
张学文：“字符多项式与表格数学”[J]《潜科学》第 39 期 2005，（1）
邹晓辉：协同智能计算知识数据库的设计方法[J]潜科学（第 39 期）2005（1）
邹晓辉：重构“概念分类体系”的新思路与新方法——从“语义三角”到“语法关系”再到“语义三棱”[A]第六届(国际)汉语词汇语义学研讨会论文集[C]
邹晓辉：优化“语义信息处理”的新方法与实施例——从“一词泛读”到“释义字组”再到“一字精读”[A]第六届(国际)汉语词汇语义学研讨会论文集[C]
邹晓辉：解析“字与字组的关系”探索“汉语形式化”新路[J]潜科学（第 41 期）2005，（3）

全国第八届计算语言学联合学术会议(JSCL-2005)交流论文

“默契通信”与“间接计算”对“自然语言处理”的重要性

——由“个性化前台”与“标准化后台”支持的“理解”

摘要：基于“文本总量控制模型”和“音节总量控制模型”的人与机高度协作且优势互补，突出“默契通信”和“间接计算”对“自然语言处理”的重要性。

关键词：自然语言理解、默契通信、间接计算、间接形式化、中文信息处理

绪言

本文涉及：“自然语言处理”的方式及习惯做法能否优化的根本性问题，属于：计算语言学的理论基础研究领域。

其特点是：以“默契通信”和“间接计算”以及“间接形式化”方略，尝试 “自然语言处理”（本文以“中文信息处理”为例）的优化。

其重点是：本文所述的原理和方略，以自然人与计算机的明确分工、有机协作、优势互补为前提。

研究途径：从人机分工和中英文对比的角度，探讨“中文信息处理新方法”（汉语的“间接形式化”方法）依据的“默契通信”和“间接计算”原理及其对“自然语言处理”的重要性。

其局限是：实施“默契通信”和“间接计算”的“间接形式化”系统——基于“两表”的“协同智能计算系统”（以“中文信息处理”为例），需有共同的“基准参照系”（“理论上可演绎”）与“应对参照系”（“实践上可相对完全归纳”）—— “目标域”由用户“约定”。

基本假设：对“自然语言”而言，“基准参照系”和“应对参照系”虽是隐性的分散存在的事实，但不是显性的工程整合的事实，所以，需把同一语种（如：中文）非线性的“信息处理方式”转化为“自然数字”线性的“信息处理方式”，从而，把“个性化前台”与“标准化后台”结合提供一种基于“两表”的“默契通信”和“间接计算”——“有针对性的重用”或“理解”自然语言。即：通过“间接形式化”，实现“对象、概念、符号、关系”的“分合切换”——针对“用户的具体重用要求”而“重构或重组”。

知识贡献：明确“默契通信”原理和“间接计算”原理以及“间接形式化”方略及其对“自然语言处理”的重要性。

希望本文能给学界同仁有所助益或启示！同时，希望听取各方面的意见或建议！

综述

下面，先从一个独特的角度，探讨“通信与计算”的相关问题。接着，探讨中英文信息处理的区别和联系。最后，指出关键问题。

0、“通信与计算”的相关问题

一方面，“通信的数学模型”（哈特莱和申农）似乎没有进一步的改变——仍然限于“超越方程”：

（1）1928哈特莱“把信息理解为（在通信符号表中）选择通信符号的方式，即：“S”（符号表中符号的个数）的“N”（被选符号序列的长度)次方，并用“选择的自由度”来计算“信息量的大小”，进而提出了信息量公式：H = N log S。

（2）1948申农“在进行信息的定量计算的时候明确地把信息量定义为随机不定性程度的减少”，进而提出了的信息量公式：$Hs(p_1, \ldots, p_n) = -K \sum p_i \log p_i$。

另一方面，“计算的数学模型”（图林等人）似乎也没有进一步的改变——仍然限于“英语思维”：

（3）上个世纪30年代，英国数学家Turing（图灵）为破解德国发明的密码机所产生的密电码，通过研究“可计算数” 或“可计算性”，即：研究能在有限的机械步骤内产生的密码，从而引出了通用计算机的概念。

（4）John Von Nouma（冯·诺依曼，美藉匈牙利人，1903－1957）通过研究“计算与存储的关系”提出“通用计算机的体系结构”——著名的"冯·诺依曼机"设想，其中心就是有存储程序原则——指令和数据一起存储．这个概念被誉为'计算机发展史上的一个里程碑"，它标志着电子计算机时代的真正开始，指导着以后的计算机设计。

（5）1949年，J.Mauchly提出了短指令码的概念，即让编程人员用熟悉的加、减、乘、除等符号编写程序，通过一个预制的表格将这些符号变成“短码”，再变成机器码。“短码”概念的引入使人们跳出了机器码的约束，“短码”（汇编语言）也由此成为一切高级程序语言最重要的基础。

国际通用的数字计算机及其通信（形式信息）交换标准[如：美国标准信息交换码（ASCII）]的确立[注：能表示几乎世界上所有书写语言的字符编码标准（Unicode）也是在这个思路的基础之上发展或扩充的]，哈特莱-申农贡献的“（形式）信息概念”具有科学的奠基作用。邱奇=图灵“可计算数”，"冯·诺依曼机"设想，J.Mauchly“短码”以及在此基础上以“数学、逻辑、语言”乃至“图形”等“人工语言的形式体系”为特点的“高级程序语言”，都是基于“英语字母”和“英语思维”的方式而逐步发展起来的。可以说，与“汉语笔画”和“汉语思维”的方式“几乎无缘”。

1、区别探源

众所周知，英文是“小字符集”，中文是“大字符集”，这是造成“英文信息处理”与“中文信息处理”之间一系列区别的一个根本原因。

仅仅限于用计算机科学领域的所谓术语这么一说，一般的专业人员通常也难以直观地想象出“英文信息处理”与“中文信息处理”之间的实体区别或本质差距。不熟悉“英文打字机及其构造原理”与“中文打字机及其构造原理”之间具体区别的人，对此自然“不知所云”。

事实证明：通用计算机的“键盘、字库、字处理标准”的发明几乎统统渊源于“英文打字机及其构造原理”——与“中文打字机及其构造原理”完全是“风马牛不相及”。计算机出现以前，人们是怎样借助英文或中文打字机进行英文或中文“信息处理”或“通信与计算”的呢？了解这个问题答案的读者，对此也许应该会有较为深入的感受和理解。

2、联系探源

我们知道，数字电路的“开、关”， 神经细胞的“ 兴奋、抑制”， 二进制数的“0、1”之间的联系或一致性，这是带来“计算机信息处理”与“自然人信息处理”之间历经一系列“中介”

或“输入输出转换”之后可能联系的一个根本原因。

仅仅用计算机科学、数学、生理学领域的所谓行话这么一说，一般的专业人员通常也难以直观地想象出“计算机信息处理”与“自然人信息处理”之间的实体关联或本质联系。不研究“计算机与自然人”如何具体分工协作的人（既不熟悉计算机电路，又不熟悉自然人神经系统，难以理解其中的道理），对此自然“云山雾罩”。

事实证明：从计算机发明之初到广泛普及的今天，凡是与计算机相关的领域——无论是硬件、还是软件、甚至是数据库，几乎统统涉及“计算机与自然人之间”如何具体分工协作的问题（从事计算机研究或应用计算机工作的具体人员，虽不一定都能明确地认识到这个问题，但这的确是一个基本事实）。最典型的例证，一是各种样式的“计算机辅助”产品——这是典型的“机助人”现象；二是各种类型的“软件工程”服务——这是典型的“人助机”现象。

3、关键问题

区别与联系的关键涉及：如何在计算机与自然人的“基本输入输出系统（BIOS）”之间建立一系列“中介”或“相互传递或转换”（“通信与计算”）的函数关系？本文关心：中文与英文在“键盘、字库、字处理标准”上有无“兼容”的可能？基于“字”的“中文信息处理”与基于“词”的“英文信息处理”在“通信与计算”模型上有无“实质关系”？

换一句话说，是否存在可能改变以下格局的原理及方法？

当前的格局：现在通用的所谓汉化的“中文信息处理”统统是建立在“基于英文的计算机底层技术”的基础之上的。可以说，到目前为止，还没有“基于中文的计算机底层技术”。

有谁见过哪一台通用计算机“一开机”其“基本输入输出系统（BIOS）”就能直接显示中文？又有谁见过在哪一台通用计算机“裸机”上直接采用“基于中文的编程语言”？至今没有。如何从根本上改变这种格局？见：方法、结果、结论和议论。

方法

0、“人机协作”

“人机合作”与“人机竞争”相反相成。“人类智能”与“人工智能”的关系，涉及“竞争”与“合作”两种基本形式。本文提倡“人机合作”或“人机协作”——以“（自然人与计算机的）明确分工、有机协作、优势互补”为前提。基本步骤：（1）“分工”：“静态信息处理”以“人助机”为主；“动态信息处理” 以“机助人”为主。（2）“ 协作”：“首次处理”从“专家用户”那里“获取”知识或信息；“再次处理”从“大众用户”那里“获取”知识或信息。（3）“ 互补”：“首次重用”由“标准化”系统“表达”知识或信息；“再次重用”由“个性化”系统“表达”知识或信息。具体操作｛ 除“三多”（多媒体、多语种、多学科）之外，见：本文的姊妹篇——“中文信息处理新方法”[(JSCL-2005)论文]｝，涉及：“两表”、“三多”、“三化”、“三注”。

1、“间接形式化”

“间接形式化”与“直接形式化”相反相成。这里以“汉语的间接形式化”为例，基于“两表”的“三多”实现“对数字与字及字组（乃至：语音）的直接呈现”。旨在：变“不可计算”为“可计算”。其中，前“两多”可“间接计算”；后“一多”可“默契通信”。

2、“间接计算”

“间接计算”与“直接计算”相反相成。这里以“汉语的间接计算”为例，基于“两表”的“三化”实现“对数字的直接计算与字及字组（乃至：语音）的间接计算”。旨在：变“难计算”为“易计算”。

3、“默契通信”

"默契通信"是一种令人满意且相当理想的"通信"。所谓"默契通信"，简单说，就是：在通信方法上“心照不宣”，在通信效果上“不谋而合”。日常生活中几乎人人都有过"默契"的快感。我们试图把各种令人愉快的"默契方式"推广到“计算机辅助”和“互联网辅助”领域。基本方法：基于“两表”的“三注”实现“对（表达知识或信息的）数字或字及字组（乃至：语音）有针对性的

重用”。旨在：“去冗传要”—— 显著地减小“知识重传”的次数，即：只需传“本真信息”而不必“重传”大量的“重用知识”及相应的“形式信息”。在此，“本真信息”就是如何在“两表”中“有针对性地重用知识或信息”的“序位信息”。“形式信息”（如：“音像或符号信息”）与“内容信息”（如：“知识信息”）相反相成。

有关注意事项

（1）关于“合作纽带”：不仅具有“强人工智能”与“弱人工智能”的“计算机”及“互联网”而且具有“强智”与“弱智”的“自然人”均可“分类分批”纳入“人机协作”的“合作机制”中来重新规划。这里强调：提供跨领域的“标准化后台”与“个性化前台”的计算机辅助和支持，为所有参与到“间接计算”和"默契通信"的“间接形式化”的“合作系统”中来的各方提供一个共同的“基准参照系”以及由此“可演绎”且受“相对完全归纳”、“枚举”或“类比”约束的一系列“应对参照系”。（2）“基准参照系”：Z_ASCII（中国标准信息交换码）或GTCM（文本总量控制模型）的“0”分表（见：“中文信息处理新方法”）。（3）“应对参照系”：GTCM的“1-12”分表（见：“中文信息处理新方法”），其中的“数据、信息、知识”的“处理方式”涉及：“目标域”（对具体的用户而言，涉及：“两表”的“未知域”与“已知域”）。（4）“三化”：在GTCM的“1-12”分表（见：“协同智能计算语言数据库”）与GSCM（音节总量控制模型）的“1-多（字组）”分表（见：“字与字组的关系”的解析）中体现的“定义表格化、字组数字化、义项字组化”。（5）“三注”：GTCM的“4-6”分表；GSCM的“1-多（字组）”分表（见：“中文信息处理新方法”）。（6）“三多”：GTCM的“0-12”分表（见“协同智能计算知识数据库”）。

结果

基于“两表”的“间接形式化、间接计算和默契通信”，既可为“用户”提供跨领域“自然语言理解”的“计算机辅助和支持”功效，也能为进一步优化“协同智能计算系统”创设更为有利的条件。

0、判断是否“理解”自然语言的标准

“协同智能计算系统”能否做到“有针对性地重用”相应的“知识信息数据”？“ 能”则视为：具有“理解”能力，“ 否” 则视为：不具有“理解”能力。

判断“是”与“否”的依据是基于“两表”的“选域定向、测序定位”，相当于“经纬指南、街区门牌”或“街区门牌、对号入座”的作用。其中，“选域定向、测序定位”通过“后台程序的计算”实现；“街区门牌、对号入座”通过“前台界面的呈现”实现。具体分工协作如下：

1、“理解”与“间接形式化”的关系

“有针对性地重用”，可区分为：“后台重用”与“前台重用”。基于“两表”的"间接形式化"的“两列”或“多列”由于“相互对应”，所以，只需“前台呈现”即可达到“形式信息处理”的目的。这就为“协同智能计算系统”提供了“标准化后台与个性化前台”融合的“计算机辅助和支持”环境。

2、“理解”与“间接计算”的关系

“有针对性地重用”，可区分为：“直接重用”与“间接重用”。基于“两表”的"间接计算"的“两列”或“多列”由于“相互对应”，所以，只需“间接重用”即可达到“数据计算”的目的。这就为“协同智能计算系统”用户应用系统如何简化“数据结构”和“算法”提供了条件。

3、“理解”与“默契通信”的关系

“有针对性地重用”，可区分为：“全盘重用”与“关键重用”。基于“两表”的"默契通信"的“双方”或“多方”由于“理解到位”，所以，只需“关键重用”即可达到“信息交换”的目的。这就为“协同智能计算系统”用户之间的“信息”或“数据”的“优化传输”创造了条件。

结论

基于“两表”的“默契通信”和“间接计算”对“自然语言处理”十分重要。

1、“间接形式化”可解决“自然语言处理”操作难的问题

“间接形式化”，有利于：“自然语言处理”在“操作”上的“人机分工——扬长避短”。

实验证明：可有效地解决“中文信息处理”各个层次的“歧义消解”问题，进一步拓展还可解决“自然语言处理”涉及的“识别、理解、表达”等“人工智能”问题。

好处：“静态信息处理”可“共享”易“重用”。“动态信息处理”易“操作”（需后续步骤配合）。

例如：汉语的“字内、字间、字外”信息处理，涉及“文字、语音、词汇、修辞、语义、语法、章法、逻辑、语用、翻译”等具体的汉语及中文信息处理。

2、“间接计算”可解决“自然语言处理”计算难的问题

“间接计算”，有利于：“自然语言处理”在“计算”上的“人机协作——相互融合”。

实验证明：可有效地解决“中文信息处理”各个层次的“信息计算”问题。

好处：“人助机”＋“机助人”＝“个性化前台”与“标准化后台”在“计算”上的融合。

例如：对汉语的词汇信息处理，可通过GSCM的“1-多（字组）”分表，实现“人机互动”高效“切分”和“标注”。

3、“默契通信”可解决“自然语言处理”质量低的问题

“默契通信”，有利于：“自然语言处理”在“质量”上的“人机互补——取长补短”。

实验证明：可有效地解决“中文信息处理”各个层次的“信息交换”问题。

好处：“大协作”＋“好算法”＝“个性化前台”与“标准化后台”在“质量”上的互补。

例如：汉语的文本信息处理，可通过GTCM的“1-12”分表，实现“人机互动”高效“处理”。

总而言之，″间接形式化″的关键是“标准化后台与个性化前台”融合“优化操作”。“间接计算”的关键是“优化算法”。“默契通信”的关键是“优化传输或信息交换”。

总体来说，基于“两表”的“人机之间的高度协作且优势互补”可较方便地解决“自然语言处理”的上述“关键难题”或“瓶颈问题”。

议论

从GTCM与GSCM两个总量控制模型各个分表的自然语言处理的实际情况来看，“默契通信”和“间接计算”对“自然语言处理”的重要性，尤其值得“中文信息处理”学界的关注或重视。因为，其中涉及一系列有待进一步探讨的具体课题。现在举例提示如下（希望有兴趣的学者参与探讨！）：

1、就汉语的文本信息处理而言，通过GTCM的“0-12”分表，我们至少发现了以下极有潜力的课题

（1）GTCM的“0”分表，涉及的Z_ASCII（中国标准信息交换码）体系设计或确认的问题；（2）GTCM的“0，1，2，3，4”分表，涉及的“汉语形式体系”中“广义字本位与狭义字本位的关系”和“字内信息处理”的问题；（3）GTCM的“4，5，6”分表，涉及的“（汉语的）字、辞、块”与“（英语的）词、词组、短语”的关系问题；（4）GTCM的“4，5，6，7，8”分表，涉及下述的“十大（微）系统工程”；（5）GTCM的“9，10”分表，增加“章法信息处理”；（6）GTCM的“11，12”分表，增加“分类与目录信息处理”。

2、就汉语的音节信息处理而言，通过GSCM的“1-多（字组）”分表，还涉及以下极有潜力的课题

（1）通过“一字和二字的关系”的解析（既基础又典型的“字间信息处理”）和“文字、语音、字典、词汇、修辞、语义、语法、逻辑、语用、翻译”等具体学科涉及的基于GSCM的“1-2（字组）”分表的“汉语及中文信息处理”的“十大（微）系统工程”；（2）在（1）的基础之上，向GSCM的“3-多（字组）”分表的逐级延伸的“组合与分解”（“字间信息处理”）；（3）在（1）和（2）的基础之上，由“无标点符号”（仅限于GTCM的“4，5，6”分表）向“有标点符号”（拓展到GTCM的“4，5，6，7，8”分表）延伸的“组合与分解”（由“字间信息处理”到“字外信息处理”）。随着（1）、（2）、（3）或GTCM的“4，5，6，7，8”分表的逐级延伸或拓展，“十大（微）系统工程”的工作任务

也随之相应地充实或丰富。

3、几个特殊的中文信息处理课题

（1）“字处理”由“狭义”向“广义”拓展过程中涉及的“汉语感知方式”与“中文信息处理”的关系问题（2）“词处理”由“静态对比”向“动态对比”过程中涉及的“汉语感知方式”与“英语感知方式”以及“中文信息处理”与“英文信息处理”的关系问题（3）“句处理”由“语义、语法、语用”向“文字、语音、词汇、修辞、语义、语法、章法、逻辑、语用、翻译”拓展过程中涉及的“汉语感知方式”、“汉语思维方式”与“中文信息处理”的关系问题（与“词处理”和“篇章处理”相结合）（4）“篇章处理”与“图书信息处理”由“简单的分类与编目”向“结构化、标准化、数字化”的方向发展，如：把“分类目录的信息处理”与“字、词、句的信息处理”结合乃至贯穿GTCM的“0-12”分表的信息处理。

*WordNet-online version；*OpenCyc.org；*CLSW1-6；*ICCC2005* 5th IEEE-ACM International Workshop on Grid Computing (Grid 2004)；

*张钹院士（2003年8月8日光明日报）谈：基础研究对于技术创新的重要性

*孙茂松教授谈：中文信息处理领域面临的机遇和挑战

参考文献（按照公开发表的时间先后排序）

R.V.L.Hartley（哈特莱）.1928,Transmission of Information,BSTJ,Vol.7,p.535-536.

Church, A（邱奇）, 1932, A set of Postulates for the Foundation of Logic, Annals of Mathematics, second series, 33, 346-366.1936, A Note on the Entscheidungsproblem, Journal of Symbolic Logic, 1, 40-41.

Turing, A.M（图灵）, 1936, On Computable Numbers, with an Application to the Entscheidungsproblem, Proceedings of the London Mathematical Society, Series 2, 42 (1936-37), 230-265.

C.E.Shannon（申农）.1948, Mathematical Theory of Communication,BSTJ,Vol.27,p.379-423,632-656.

陈肇雄主编：机器翻译研究进展[C] 1-564 页，电子工业出版社，1992

徐通锵：语言论--语义型语言的结构原理和研究方法[M] 295-442 页，东北师范大学出版社， 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] 1-516 页，清华大学出版社，1998

林杏光：词汇语言学和计算语言学[M60-118，140-376，语文出版社年，1999

俞士汶、朱学锋：计算语言学文集[C] 1-254 页，北京大学计算语言学研究所，2000

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

鲁　川：汉语语法的意合网络[M]1-277 页，商务印书馆，2001

邹晓辉：一种知识信息数据处理方法及产品[J]发明专利公报 G06F163 知识产权出版社，2000，（11）

俞士汶：关于汉语信息处理的认识及其研究方略[J]语言文字应用（总第 42 期）2002，（2）

苏培成等编：语文现代化论文集[C] 商务印书馆，2002

黄河燕主编：机器翻译研究进展[C] 1-282 页，电子工业出版社，2002

徐波、孙茂松、靳光瑾主编：中文信息处理若干重要问题[C]科学出版社出版，2003

邹晓辉：协同智能计算语言数据库的设计方法[J]潜科学（第 32 期）2004（7）

邹晓辉：协同智能计算知识数据库的设计方法[J]潜科学（第 39 期）2005（1）

全国（北京2005）信息科学交叉研究学术研讨会论文

语义信息新论

——推定信息科学的基本公式

（注意逻辑语义与词汇语义的区别）

摘要

有人说：语义信息的定义是不清楚的。与之不同，语义信息新论建立在哈特莱-申农探讨的信

息与其他人探讨的语义信息的基础之上。依据融智学的信息理论，任何一个数据库中存储的数据都可分为信息与知识两部分，哈特莱-申农探讨的信息可归结为数据，其他人探讨的语义信息可归结为知识。因此，可推定信息科学的基本公式：I = D - K，意思是：狭义的信息（I）是数据（D）中取出知识（K）的余留部分，也是将获得的那部分知识；广义的信息包含：数据、狭义的信息、知识。依据语义学研究经验，学界应重视：被语言学家和逻辑学家，尤其是近期被计算语言学者，密切关注的词汇语义和逻辑语义的基本语义问题。因此，可借助该基本公式重新讨论语义信息。

关键词：语义、信息、数据、知识

一、绪言

现有的语义信息定义是不清楚的。与之不同，语义信息新论主要论述：语义（注意：逻辑语义与词汇语义的区别）、信息、语义信息以及信息科学的基本公式，**研究领域**涉及信息科学的基本理论问题。**特殊性**：从信息、数据、知识三者关系的角度，深入探讨信息的本质和信息科学的基本原理。**重要性**：明确提出信息科学的基本公式和其中蕴涵的狭义信息与广义信息以及语义信息等概念。强调区分逻辑语义与词汇语义对具体探讨语义、信息、语义信息的作用。**研究途径**：在哈特莱-申农探讨的信息与其他人探讨的语义信息的基础之上，依据融智学的信息理论——“语义三棱”模型，会发现：哈特莱-申农探讨的信息可归结为数据，其他人探讨的语义信息可归结为知识。这支持了“数据=信息+知识”（D = I + K）的基本假设，进而可推定信息科学的基本公式和给出一般科学的信息定义以及明确信息科学的基本原理。结合逻辑语义与词汇语义的区别可搞清楚最基本的语义问题。借助上述研究成果可重新讨论语义信息，通过追根寻源可探寻逻辑语义的自然人判定与逻辑歧义的计算机处理的理论，通过旁征博引可考察词汇语义的自然人判定与词汇歧义的计算机处理的实践。**局限性**：由于二歧性是多歧性通向无歧性或唯一性的必经环节，又由于词汇语义是句子、语篇乃至语境的语义问题研究的基础，所以，本文仅以逻辑语义的二歧性与词汇语义的多歧性为例，考察逻辑语义与词汇语义的区别，以便对语义、信息、语义信息等概念进行严格定义。本文注重对狭义信息的探讨，仅指出进一步探讨广义信息的途径。**基本假设**：任何一个数据库中存储的数据都含有信息与知识两部分，哈特莱-申农探讨的信息可归结为数据，其他人探讨的语义信息可归结为知识。数字计算机是基于逻辑语义而建立的信息处理装置，就逻辑推理和数学计算而论，其基础旨在：消除逻辑歧义（即：消除路径选择的二歧性）和实施二进制计算（注：其它进制均可与二进制之间实现自动转换）；自然人是基于词汇语义而建立的信息处理装置，就自然语言理解而论，其基础旨在：消除词汇歧义（如：消除义项选择的多歧性）；可把逻辑语义视为词汇语义的特例，即：把逻辑语义与词汇语义的区别视为二歧性与多歧性的区别。**知识贡献**：1、明确提出信息科学的基本公式 I = D – K 即：（狭义）信息 = 数据 - 知识。2、明确界定信息科学的核心概念和研究对象，即：（狭义）信息是数据中取出知识的余留部分；（广义）信息包含：数据、（狭义）信息、知识。3、明确提出信息科学的研究策略和基本方法及研究途径，即：重点研究（狭义）信息与整体把握（广义）信息以及从数据和知识两端逼近（狭义）信息。4、建议采用协同智能计算的方式，把语言学家、逻辑学家、计算语言学者以及具体应用领域用户密切关注的逻辑语义与词汇语义的基本歧义问题，进行系统化处理且不断优化。

二、综述

1、信息概念有狭义与广义之分，相应的科学理论也如此

狭义的信息科学，通常指：通信与计算机科学，主要涉及：信息的形式化概念、原理及处理方法，现已相当成熟，相应的技术与应用也十分普及。最具说服力的论据就是“计算机”及其“互联网”带来的“信息革命”——严格地讲是“形式信息革命”（注：这是融智学的观点之一）。

广义的信息科学，含：狭义部分以及其它部分，主要涉及：信息的形式与内容（涉及：“语义信息”这个十分重要而又非常含混的复合概念）的关系。众所周知，信息是信息科学的核心概念和研究对象，语义是研究的重点和难点。严格地讲：语义信息是不清楚的或欠明了的和主观的或个人的

（Semantic information is ill-defined and subjective）。因此，我们一直认为：继“形式信息革命”之后，必然会发生“语义信息革命”（注：这是融智学的观点之二）。

事实证明：与狭义部分相比，广义部分显得还很不成熟，其具体研究领域也显得十分凌乱。例如：申农“信息论”之后，人们虽然提出了各种各样论及“语义信息”或“广义信息”的理论，但至今无一获得“形式信息”理论那样的一致公认。这说明“语义信息革命”仍在襁褓之中。

所谓“新论”就是针对上述“旧论”而提出来的。其基础和重点是论述信息科学的核心概念、基本原理、基本公式，难点是语义分析——关键是区分逻辑语义与词汇语义，两者虽然有关但是不能混淆。

2、“信息是 …”与“…（的）信息”

信息是什么？都有哪些基本类型的信息？目前为止就是自然人、专家们几乎也都“说不清，道不明”，又如何强求计算机系统能消解其中的各种歧义呢（“中文信息处理”和“基于汉语的知识表达”的例子随处可见）？

为此，本文特意提炼出“信息是…”与“…（的）信息”两个典型表达式作为信息概念研究的重中之重，着重从逻辑语义与词汇语义的分析，试图对信息（Information）、语义（Semantic）、语义信息（Semantic Information）进行一番论证。旨在深入探讨信息的本质和基本类型或分类的问题。

本文（《语义信息新论》）着重对狭义信息的微观分析，其姊妹篇（《广义文本与本真信息》）着重对广义信息的宏观分析。希望对学界能发挥“抛砖引玉”的作用（如有“牵一发而动全身”的效果，将是最理想的！）。

3、识别、理解、表达的消歧难题

上述两个典型表达式，不仅信息一词之后“…”省略的说明存在“随机不定性——不确定”，而且信息一词之前“…”省略的说明也有多种可能的选择方式——存在“多歧性”。

如何消解上述“不确定”或“多歧性”？一直都是自然人的“逻辑思维”所面临的“语义”问题与计算机的“符号识别、语言理解、知识表达”所面临的“语义”问题。实质是如何消除歧义的问题。

古代的柏拉图与亚里士多德两位哲学家都曾被“歧义”难题所困扰。

近代的弗雷格与布尔两位数学家及逻辑学家分别以不同方式提出了各自解决逻辑歧义难题的形式体系。弗雷格还试图发现解决词汇歧义难题的形式化问题 [语义上的形式化研究的其它路径还有：晚些时候索绪尔的普通语言学的关系（涉及：词法、句法、…乃至整个语言系统的语义关系）研究以及其后乔母斯基的形式语法研究]。近代哲学领域发生的“语言转向”就始于弗雷格对语义问题的探讨（以后的“语义三角”和“意义理论”的探讨也都渊源于此）。

这之后，对语义问题或歧义难题的思考虽从未间断，但几乎就再没见过（具有与之等量齐观的重大研究成果或）根本性突破（目前哲学正处于“信息转向”的初期）公诸于世。

尽管如此，计算机科学、认知科学（含：人工智能与认知心理学）、计算语言学（研究：自然语言处理及中文信息处理）等交叉学科各具体研究领域，还接二连三地公布了不少重要研究成果（或大大小小的突破），如：基于规则、基于统计、基于实例及其相互结合的各种消歧原理及方法，有些仍在屡屡翻新（见：人工智能与计算语言学研究领域，其中，自然语言的词汇语义处理是最基础的）。

基于前人和他人上述广泛的研究，工程融智学及其典型实例公开的消歧原理及其“两表、“三化、三注、三多”的（人机）协同智能计算系统的一系列设计方案，才可望问鼎这个跨越多个世纪而令人生畏的语义上如何系统化消歧的难题，并首先在理论思考上获得根本性突破。

可以说，人们对信息的各种认识分歧皆可归因于上述语义上的消歧难题。反之，如能在逻辑语义与词汇语义这一基础层面系统地解决语义上的消歧难题，也就可在根上解决信息的一般科学定义和系统分类的问题——消解人们对信息的认识分歧。

搞清逻辑语义与词汇语义的区别很重要。在基础理论上，这对确定信息本质及分类，作用明显。

4、以往对信息、语义、语义信息等概念的探索

通信科学界公知的信息概念由哈特莱（1928）提出，其信息量计算公式由申农（1948）在数学上发展成了（经典）信息论，同期，确立了（数字化通信）信息量的基本计量单位 bit。

计算机科学界公知的美国标准信息交换码（ASCII），明确了信息科学技术的“形式化”基础。

以上关于信息的描述实质上可归属于数据范畴；其它各种关于信息的认识（如：钟义信《信息科学原理》所转述的其他几十种关于信息的定义、说法、解释、说明、…）几乎都属于知识的范畴。

《一种知识信息数据处理方法及产品（发明）》（2000）和《融智学（新范式）》（2000），明确提出了“义、文、物、意”融智概念体系（通论）、信息基本法则（通则）和多元数表达式（通式）。《协同智能计算语言数据库的设计方法》（2002）明确给出了“子全域”、“超子域”及其“进阶层式”概念、原理与方法的典型实例。《协同智能计算知识数据库的设计方法》（2002）明确给出了“已知域”、“未知域”及“目标域”概念、原理与方法的典型实例。《融智学纲要》（2002）明确区分了（广义融智学的）哲学信息观与（狭义融智学的）科学信息观。在信息科学理论和信息技术实践的有关科学论文中开展了一系列具体研究（2003-2005）。这些就是以下确定信息科学的核心概念、基本原理、基本公式的思想渊源。

以下试图就信息及语义信息的本质与基本类型或分类提出一种新观点及新方法。

三、方法

语义信息新论对语义、信息、语义信息的界定方法及步骤：

0、设定前提

探讨语义，限定在逻辑与词汇两个层面（涉及：路径消歧与义项消歧）。

探讨信息，限定在科学范围（涉及：计算机科学的可计算性与数据处理）。

探讨语义信息，限定在信息科学范围（涉及：信息处理以及知识处理）。

1、界定语义

限定论域，即：只讨论词汇语义与逻辑语义及其相互关系。

语义，主要指：词语意义。句子、段落、篇章、语境的语义皆可由上述两种类型的语义推演而知。

语义内涵及其重要性

“（融智学）语义三棱”比“（语义学）语义三角”有更加丰富的语义内涵。特点：揭示意与义的关系。如果说“意义问题是当今人文科学（含：哲学）研究的核心问题”，那么，意与义的区分，则是（整个）科学（含：人文科学、自然科学、人工科学、……）研究的核心问题。

逻辑语义与逻辑歧义消解

解决逻辑语义的二歧性问题：“ … 是 … 还是 … ”？可用二值逻辑消解逻辑歧义的基本表达式：“ … 是 …而非 … ”。例如：逻辑语义涉及的歧义是逻辑歧义还是词汇歧义？答案显然是前者。又如：词汇语义涉及的歧义是逻辑歧义还是词汇歧义？答案显然是后者。可见逻辑语义与词汇语义关系密切。

词汇语义与词汇歧义消解

解决词汇（含：字汇——汉语的特点）语义的多歧性问题：如何选择“（…）义”？

一般表达式：“… （的）… ”[注：（的）前省略（…）填“用字”——用以限定“解字”的义项搭配的字；（的）后省略（…）填“解字”。只有“用字”与“解字”之间不能接续时，才需插入（的）字而构成三字组——多字组另议]。

需要添加“用字”——如：语、主、含、本，才能限定待解释其义项的字——“解字”，如：义。仅仅把“语”与“义”两字结合构成“语义”这个二字组还不足以判定它究竟是逻辑语义还是词汇语义，必须延长字组才能消歧。借助字组表筛选“用字”并查询细分知识领域，消除“解字”义项的词汇歧义。

2、界定信息

提出数据公式，明确界定范围。

数据公式

“ D = I + K ” [“ Data = Information + Knowledge”的缩写，汉语意思是：数据 = 信息 + 知识]

公式中，“数据”集合，为“目标域”（“限定范围”如：数据库），其特征是“可计算”；“信息”集合，为“未知域”，其特征是“可选择”；“（各门科学）知识”集合，为“已知域”，其特征是“可重用”。

说明：（数据 Data）**D =** （信息 Information）**I +** （知识 Knowledge）**K**

（全部数据）**目标域 =** （信息或未知数据）**未知域 +** （知识或已知数据）**已知域**

（全部数据）**目标域 =** （未知部分的数据）**未知域 +** （已知部分的数据）**已知域**

推出信息公式，揭示信息本质。

信息公式（这是一般科学的信息定义式，含：狭义信息与广义信息两个重要方面的概念）

“ I = D - K ”[“Information = Data – Knowledge”的缩写，汉语意思是：信息 = 数据 - 知识]

信息定义

（作为信息科学的核心概念和研究对象的）（狭义）信息是数据减去知识的余下部分，其形式特点是还未承载知识的数据；其内容特点是将获得的知识。（广义）信息包含：数据、（狭义）信息、知识。

信息本质

（狭义）信息与知识的区别在于：信息是未知的，知识是已知的。（狭义）信息与数据的区别在于：信息是还未承载知识的那部分数据（知识是承载知识的那部分数据）。信息的形式与数据有关，信息的内容与知识有关。（广义）信息的说法过于粗放而不严谨。（狭义）信息才是严格的科学定义，其本质是数据的序位，其中未知与已知两部分的划分仅仅是相对于主体（如：人）或载体（如：信源、信道、信宿）而言的。

推出知识公式，揭示语义信息

知识公式

“ K = D – I ”[“ Knowledge = Data - Information”的缩写，汉语意思是：知识 = 数据 - 信息]

3、界定语义信息

语义信息的说法也过于粗放而不严谨。人们在说“语义信息”时，实际上是强调（狭义）信息的内容特点——“将获得的知识”。

一旦认清了语义的内涵和信息的本质，具体界定语义信息的方法也就明确了。即：根据语义三棱，从宏观上深入地解析“语（广义文本）”与“义（本真信息） 两方面（本文的姊妹篇《广义文本与本真信息》将对之做深入的探讨）；根据信息公式，从微观上优化数据处理与知识处理进而可系统地解析语义信息——（狭义）信息的内容特点——“将获得的知识”。

四、结果及结论

上述方法产生的必然结果和有益效果以及相应的结论

确立信息科学的基本公式，可导致如下基本结果及结论

明确一般科学的信息公式，为统一信息概念提供了可计算可测量的数学模型

由于一般科学的信息公式——信息科学的基本公式的确立，（狭义）信息与（广义）信息的关系得以明确。这不仅可消解人们在信息概念问题上长期存在众多认识分歧（这种现象实际上是在认识上的“盲人摸象”，也是理论不成熟的必然表现），而且，可解决一般科学的（狭义）信息理论与（广义）信息理论以及（信息）科学与（信息）哲学之间长期缺乏界定标准的问题。进而，明确信息科学各个学科的研究对象。

对信息科学建设的推动作用

理论上，可明确信息科学的三大基本原理——信息与数据的关系原理；信息与知识的关系原理；数据与知识的关系原理。

实践上，可总结以下基本认识——以往对信息的探讨，由于对信息、数据与知识三者之间的关系缺乏清晰认识，常常张冠李戴；以往对数据与知识及其关系的探讨，为深入探讨信息奠定了必要的基础（如间接的理论和示例）。

对以往信息概念模糊的澄清

可明确“（狭义）信息”以及探讨与区别“信息、数据、知识”的重要性；可明确“（广义）信息”以及探讨与“科学、技术、艺术、哲学”的联系。澄清（狭义）信息与（广义）信息的概念及其相互关系之后，既利于把握一般科学的“信息”本质，又利于认清具体科学的“信息”特征。

对以往算法路径的必要拓展（本文的姊妹篇《广义文本与本真信息》将对之做具体的介绍）

在认可并继承“直接基于指数方程的对数计算方法”的同时，发展出了一套“间接基于代数方程的算术计算方法”及其优化策略和计算路径。

五、总结和议论

总之，新论旨在明确上述信息科学的核心概念、基本原理、基本公式，这不仅利于信息概念的统一，而且，利于信息科学（乃至整个科学）体系的优化（本文的姊妹篇《广义文本与本真信息》将对之做必要的介绍）。

明确“语义、信息、语义信息”的其它好处（注意区别：科学研究与日常应用两方面，忽简单地混为一谈）

从逻辑与词汇两个方面排除“语义、信息、语义信息”的路径分歧和义项分歧，使复杂抽象的理论思维可且易重复操作。既可避免自然语言的歧义性对人们认知的误导，又可利用自然语言的灵活性拓展人们的认知路径和认识视野。特别是发现信息科学的基本公式之后，反过来对微观上的路径消歧和义项消歧，也有帮助或促进作用。例如：知道“语、文、字、…”与数据相通，并且，知道“义、意、意义”与知识相通，这之后，所谓“语义”就可简单而通俗地表述为“语言形式的内在含义”或“语言形式要表达的思想内容”或…。又如：知道信息与数据、知识相关之后，所谓“信息”就可简单而通俗地表述为“数据中未被理解的部分”或“数据中未知的部分”或…。因为“数据”由“未被理解的部分——信息”和“已被理解的部分——知识”构成。再如：所谓“语义信息”实质上强调的是“信息”的内容特征——准确地说是强调“数据中未被理解部分”的内容特征。

明确信息科学与（质能）科学的关系“消除”认识误区

判定科学的标准长期存在不确定性，与学界对“信息科学与（质能）科学的关系”缺乏认识（或存在认识误区）息息相关。能否“消除”这一认识误区？关键在于是否能搞清“信息科学与（质能）科学的区别和联系”。一般科学的信息公式，“在可计算数据的前提下”探讨“信息与知识的关系”，这就为学界对“信息科学与（质能）科学的关系”的界定提供“科学的判定标准”。如果说：探索未知领域的信息奥妙是科学的使命，传播已知领域的知识成就是教学的任务，那么，（质能）科学的前沿领域（涉及探索未知领域的信息）也是信息科学的一部分，信息科学与（质能）科学的知识成就都是教学内容的一个部分。

科学标准的重建

对上述结果及结论的系统解析，不仅可明确信息科学的核心概念和研究对象——巩固信息科学的基础，而且，可明确信息科学的研究任务、（创新的）研究方法和理论工具（另有融智学的系列文章系统介绍），同时，还可明确信息科学与（质能）科学的区别和联系——以利于重建（优化的）科学标准。

基本的标准体系

可计算数据的可证实性（如：具体数据被验证是在限定的目标域）与可证伪性（如：具体数据被验证不在限定的目标域），可重用知识的被认可性及被认可程度（如：不仅具体数据被验证是否在限定的目标域中，而且，该验证结果能被具体学科或学科群的科学共同体其他同行或准同行所认可——具体认可程度不仅可由认可者的人数及其在学界的实际影响或具体应用效果进行评判，而且，还可由协同智能计算系统进行自动评判

以及网络计算机辅助评判）。

具体的评判尺度

除了新颖程度；创新程度（非显而易见程度）；实用程度（简称：三度）之外，还将增加（协同智能计算系统自动评判以及网络计算机辅助评判的）可计算程度、可选择程度（算法的好坏程度）、可重用程度。

基本的评判方式（借助标准化与个性化结合的形式化描述体系的支持）

在联网计算机辅助（CA）的条件下（基于目标域限定的可计算数据）比较“（发现者提供的）可选择信息”与“（现有）可重用知识（如：常识知识库和专家知识库）”之间“三度”的差异。

参考文献

Claude E.Shannon and W.Weaver. The Mathematical Theory of Communication,The University of Illinois Press, 1963

邬焜、李琦: 哲学信息论导论[M]陕西人民出版社，1987

钟义信.信息科学原理[M].福建人民出版社，1988（北京邮电大学出版社，1996 新版 2002 再版）

Shannon,C.E Collected Papers ed.by N.J.A.Sloane and & A.D.Wyner (Los Alamos, Ca: IEEE Computer Society Press). .1993,

Simon H. A. The Sciences of the Artificial (Cambridge, Mass.: MIT Press). 1996

Losee, R. M. “A Discipline Independent Definition of Information”, Journal of the ASIS, 48.3, 254-269. 1997,

徐友渔、周国平、陈嘉映、尚杰：语言与哲学——当代英美与德法传统比较研究[M]三联书店，1996.

周斌武、张国梁：语言与现代逻辑[M]复旦大学出版社，1996

How fundamental is information? References Last updated August , 2004

Floridi, L. (forthcoming), “Is Information Meaningful Data?”, preprint available at

刘叔新：词语强制搭配的语义关系类别及其性质[A]语言学论辑[C] 1-17 页，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] 295-442 页，东北师范大学出版社， 1997

王路：世纪转折处的哲学巨匠：弗雷格[M]社会科学文献出版社，1998

Mike Beaney:1999《Meaning and Truth》[M].

K.R.Popper 原著，查汝强、邱仁宗译：科学发现的逻辑[M]沈阳出版社，1999

A.J.Greimas 原著，吴泓缈译：结构语义学研究方法[M]三联出版社，1999

俞士汶、朱学锋：计算语言学文集[C] 1-254 页，北京大学计算语言学研究所，2000

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

邹晓辉：一种知识信息数据处理方法及产品[J]发明专利公报 G06F163 知识产权出版社，2000，(11)

俞士汶：关于汉语信息处理的认识及其研究方略[J]语言文字应用(总第 42 期)2002，(2)

邹晓辉：语言及语义信息的统一参照系[J]潜科学 2002.05

黄河燕主编：机器翻译研究进展[C] 1-282 页，电子工业出版社，2002

李锡胤：关于“语义三角”之我见[J]俄语语言文学研究，第 1 期 2003

邹晓辉：义项语汇典例(SVDE)的总量控制模型(CLSW-5 paper) [J]《潜科学》第 40 期 2004(3)

邹晓辉:优化“语义信息处理”的新方法与实施例(CLSW-6 paper) [J]《潜科学》第 40 期 2004(3)

邹晓辉:解析“字与字组的关系”探索“汉语形式化”新路(专著节选) [J]《潜科学》第 41 期 2004(3)

邹晓辉:中文信息处理的新方法(JSCL-2005 paper)[J]潜科学第 42 期 2005，(5)

邹晓辉:默契通信与间接计算对自然语言处理的重要性(JSCL-2005 paper) [J]潜科学第 42 期 2005，(5)

邹晓辉：融智学应用实例 (cooperating with computer e.g.) [J]潜科学第 43 期(论文连载)2005，(5)

中国人工智能学会第十一届全国学术大会 2005 年 8 月 25 日录用论文

自然语言处理的总量控制模型——形式化标准平台

摘要 除生物基因外，美国标准信息交换码（ASCII）是最成功的代码。如改进中文处理基本单元，国际统一代码（Unicode）将会更好。本文试图提供一个逻辑和数学上相对完备而堪称终极标准信息交换码（Z-ASCII）的基因文本数据库。那时就能更好地分析和解释各个中文处理单元的含义，同时，也不仅限于支持输入、输出、交换等固有的用法。基于 Z-ASCII 的中文处理单元的新用法是最简单且最有效的。中文与英文的区别相当大，对英文信息处理系统足够的 ASCII，对中文信息处理系统却远远不够，因为，音节总量控制模型（GSCM）和文本总量控制模型（GTCM）在前者是一致的可在后者却不一致而需采用与 GTCM 相应的 GB 或 Unicode——因其太粗放而没充分顾及汉语特点，要提高中文信息处理智能化水平还需基于 Z-ASCII。

关键词 美国标准信息交换码 国际统一代码 音节总量控制模型 文本总量控制模型 终极标准信息交换码 间接形式化

1．引言

在过去几年，人工智能的研究取得了长足的进展[以中文信息处理为例：1，基于微型中文造字产生器的汉字基因芯片的产品化；2，电子辞典和计算机辅助翻译系统的产品越做越好，已有完全支持中文的计算机汇编语言（如：O 语言）]。然而也还有很多重要的问题没有得到满意的解决[3，中文信息处理的基础研究薄弱（如：汉语语言学领域“各种本位说之争”和计算语言学领域“各种资源库之战”）；4，GB 与 ASCII 之间在信息处理效率上的巨大差距仍然存在，而 Unicode 汉字处理部分几乎仍沿用 GB 的作法；5， 汉语形式化困难重重，机器翻译的消歧难题依然存在]。有鉴于此[1][2][3][4]，本文提出了一种形式化标准平台——自然语言处理（含;中文信息处理）的总量控制模型，即：音节总量控制模型（GSCM）和文本总量控制模型（GTCM）及其底层技术规范——终极标准信息交换码（Z-ASCII）[5][6][7][8]。

概述：本文属于自然语言处理与理解**领域**，涉及：机器翻译，复杂性，信息化与智能化。其应用，一方面，涉及：计算机辅助教育，如：计算机辅助汉语（英语、双语乃至多语）教学；另一方面，涉及：中文信息处理产品标准与产业发展，如：改进 GB 和 Unicode 中文信息处理基本结构单元的部分。**特殊性**：直接采用工程融智学 8 大系统工程实验的前沿科技成果[9][10]，探讨长期困扰自然语言处理与理解和机器翻译，复杂性，信息化与智能化等领域的消歧难题[11][12]。**重要性**：为解决消岐[涉及：模式识别、语言理解、知识表达（典型实例：机器翻译）]的技术瓶颈提供理论模型、计算和操作的系统工程技术方法及底层技术规范。**研究途径**：1、梳理工程融智学前期研究的有关成果，2、明确语言符号形式体系两种增长方式在计算方法上不同的技术处理特点，3、突出中文信息处理的双层结构（即：“层面型结构”与“线串型结构”），4、正式提出并强调自然语言处理（如;中文信息处理）的总量控制模型（GCM）——间接形式化（区别于直接形式化或间接而又间接的所谓直接形式化）标准平台。**局限性**：本文仅限于介绍 GSCM 与 GTCM 及 Z-ASCII 中文信息处理基本结构单元改进部分的构想或做法。其它相关内容及细节和应用实例需阅读参考文献。**基本假设**：自然语言处理的总量控制模型（GCM）——形式化标准平台，作为建立协同智能计算系统（如：融智计算机和协同智能计算网及其各种智能计算艾真体及专业化智能计算终端）的基础，其底层技术规范是 Z-ASCII，结合其基础层和中、上层技术规范一道构成间接形式化数字代码阵列（m n）。**贡献**：明确强调这种形式化标准平台——自然语言处理（如;中文信息处理）的总量控制模型（GSCM 与 GTCM 及 Z-ASCII）可把人助人的小范围默契交流推广到机助人的大范围默契通信（高效消歧）。

2．综述

2．1．对待复杂问题与几何增长乃至指数爆炸问题的策略和技术路线

中文信息处理是自然语言处理的难中之难——（对中国而言）也是重中之重。从标准信息交换码的国家标准（GB）和国际标准（Unicode）看中文信息处理存在的一个根本问题。

2．1．1．由 ASCII 到 Z-ASCII（本文仅涉及其中文信息处理部分）的标准竞争

信息产业（IT）界 Wintel（微软视窗-因特尔芯片）垄断格局形成的根基，是作为文本基因的 ASCII

（美国标准信息交换码）。就目前情况看，它不仅是英语这一自然语言处理过程中识别、理解、表达的基础，而且，也是其他民族语和程序语言乃至各种专业术语处理的基础与解释或翻译的中介。可以说，它几乎已成计算机辅助人类进行知识信息数据处理的垄断文本基因。不仅 GB 和 Unicode 都必须要与之兼容，而且基于 GB 和 Unicode 的一切软硬件也都必须要与之兼容，否则，就没有产业出路。为什么会造成今天这样（英语民族主动而非英语民族被动）的局面呢？非英语民族（如：汉语民族）有必要改变它吗？能改变它吗？如果能，那么，必须怎样做呢？

朱邦复先生提出的**汉字基因**和中文语言开发小组提出的**0 语言**从各自角度做了有益尝试，并取得了相应产品的一定市场地位。但因涉及语义信息处理这样非常复杂且常会遭遇指数爆炸而必须却又难以消歧的问题，故仅靠技术发明的浅层突破，而不从基础理论上取得实质性科学发现的深层突破，难从根上改变中文信息处理的被动局面。我们知道：朱先生所谓汉字基因实际是说概念基因（涉及：语义信息处理）。中文语言开发小组所谓 0 语言实际是基于汉语的汇编语言——程序语言汉语化、翻译或解释（涉及：语义信息处理）。由此可见，如果语义与信息的关系这一基本理论问题得不到较为满意的解决，那么，所谓汉字基因和 0 语言的技术突破也只能是散点式的阶段性突破。何况 GB 与 ASCII 之间在信息处理效率上的巨大差距仍然存在，而 Unicode 汉字处理部分几乎仍沿用 GB 的作法。显然，还须寻求理论及工程上的根本性突破。本文指出：创立 Z-ASCII（这里仅探讨中文信息处理部分）是可取之路。这是一种适应当中求变革的创新做法。

2．1．2．中文信息处理的困难——既有非常复杂的一面又有指数爆炸的一面

我们知道：就质的方面而论，语义与信息的关系以及信息、理解、智能的本质探讨，是非常复杂的问题。如：汉语语言学领域各种本位说之争，说明问题已复杂到语言学专家也都难以就一个字（如：汉语“字本位”理论所强调的“字”）的含义达成共识[13]。试问这种消歧难题如何让计算机去自动处理呢？就量的方面而论，当 n 大到计算 2 的 n 次方成为不可接受或代价不可容忍时，会出现指数爆炸问题。如：计算语言学领域各种资源库之战，说明各个研究小组之间实际上主要是在时间、精力、人力、物力、财力上拼消耗——因为各方都还没有对付指数爆炸问题的系统工程方案。上述两种情况是中文信息处理时常会遭遇而往往又捉襟见肘的。如：汉语形式化难题，特别是机器翻译的消歧难题依然存在，说明中文信息处理领域仍然沿用旧方法或继续走在以往的所谓直接而实际上是间接又间接的汉语形式化道路上[14]。显然，学界和业界都还没有找到高效解决非常复杂问题的良策和直接应用好算法的基本技术路线。近期，已有证据显示：旧观念不改变，即使再好的新理论、新方法或新途径出现，再聪明的人也会视而不见。

由于以往的对策，主要是各自为政寻求具体的形式化方法、各种程序语言和具体算法。这使学界和业界的精英们时常疲于应付。超子域进阶层式成员的特点是交叉、重叠、嵌套、复杂、非线性、几何增长。这凸显了直接处理这类问题的困难（即：自然语言处理的所谓直接形式化计算途径是死胡同！）——要么非常复杂，无从入手；要么指数爆炸，无法完成。须从根上改变被动格局。

2．1．3．我们为什么要提出间接形式化对策及方法

工程融智学和理论融智学的研究发现：实际上 GTCM 超子域进阶层式成员的问题解决，至少涉及两种等价的形式化计算途径，即：几何增长与算术增长、非线性与线性、复杂与简单。而以往的信息计算理论几乎都采用直接与前者挂钩的做法[如：哈特莱-申农的信息论，其后续者的各种计算路线也都没有走出指数-对数（仍是指数增长）模式]。本文的信息计算理论通过转换而采用后者。

根据工程融智学 8 大系统工程实验的初步结果，我们发现：GSCM 与 GTCM 及 Z-ASCII 借助数据库及数据仓库的系列电子表格及其自动编号数字代码阵列，可系统全面地实现汉语（中文信息处理）间接形式化。即：一方面，系列电子表格的自动编号可视为基于算术的自然数数字代码；另一方面，多个双列表的自动编号构成的数字代码阵列又有系统而完整且现成的数学模型支持——非常利于做进一步的自动化处理。该方法也适用于英语及其它语种乃至多媒体数据的信息处理[前提条件是工程融智学所述“字（含：数字与文字及特殊字符）、式、图、表、音、像、立体（静态虚拟）、活体（动态虚拟）”八大形式体系的广义文本基因皆可依据子全域平行层式元素异义排列序趣简美法则及超子域

进阶层式成员总量控制相对完全归纳原则和同义并列对应转换法则纳入终极标准信息交换码（Z-ASCII）文本基因构造的基准参照系和 GSCM 与 GTCM 文本进化发展的应对参照系的文化基因系统工程框架]。

研究中，我们还发现：借助 GSCM 与 GTCM 及 Z-ASCII 容易构造基于母语（如;汉语或其它语种）和算数（如; 原先不可直接计算的对象，通过双列表的系列转换，可简化为算术问题而实现间接计算）的表格化（如;由数字化阵列与对象化字组一一对应同义并列的双列表可作为中文信息处理的间接形式化标准）编程辅助语言（无论是自然语言还是人工语言乃至图形图像语言几乎都可与之建立同义并列对应转换关系）。GTCM 支持的底层技术规范 Z-ASCII 既兼容 ASCII 又兼容 GB 和 Unicode 并可使后两者得以优化。

众所周知，中文与英文的区别相当大。对英文信息处理系统足够的 ASCII，对中文信息处理系统却远远不够，因为，GSCM 和 GTCM 在前者是一致的（即：对英文信息处理系统而言 GSCM 与 GTCM 之间完全同义并列）而在后者却不一致（即：对中文信息处理系统而言 GSCM 与 GTCM 之间只有局部同义并列关系）而需采用 GTCM 的 0、1、2、3、4 进阶层式（等价于 GB 或 Unicode 中文信息单元）——GSCM 起始于 GTCM 的第 4 进阶层式。由于 GB（如：GBK）或 Unicode 中文信息单元的处理方式太粗放而没充分顾及汉语特点，因此，要提高中文信息处理智能化水平还需基于 Z-ASCII。

3．方法

首先，选域[设定：子全域、超子域、（有限）目标域、已知域、未知域，作为数据、知识、信息处理的限制范围（即：自然语言处理的前提条件）]定向。然后，测序（计算 m n）定位（即：进行自然语言处理）。

3．1．选域定向——明确五域及其相互关系

子全域、超子域、（有限）目标域、已知域、未知域，简称：**五域**。已知：**元素**个数为 n 的**集合**（如：子全域）的**子集**（如：超子域）个数为 2 的 n 次方。可知：**子全域**（Z-ASCII）元素（其对应的编号数字 n 是算术增长），**超子域**（其对应的编号数字 2 的 n 次方是几何增长），是计算类型不同的两种形式体系。以下是**化繁为简**的具体转换步骤：

（1）子全域 1-m **平行层式**，其中，每一平行层式有 1-n 个**元素**，其特点是：平行层式可列举，元素可穷举。平行层式的**实例，如**：（ASCII 中的）大写和小写的英语字母；二进制与十进制的基本数字符号；标点符号；运算符号；特殊符号；（在 Z-ASCII 中增加的）汉字基本笔画；汉语拼音符号；…[字、式、图、表、音、像、立体（静图）、活体（动像）等广义文本基因均可由此间接形式化]

（2）超子域 1-m **进阶层式**，其中，每一进阶层式有 1-n 个**成员**，其特点是：进阶层式可穷举，成员可列举。两个极端情况：**子全域**可视为 **0 进阶层式**；所有进阶层式集合可视为最大的超子域。进阶层式的**实例，如**：在汉语的 0 基本笔画、1 不成字偏旁部首、2 变形字偏旁部首、3 字中字偏旁部首、4 字、5（无虚字的）辞、6（有虚字的）块、7（标逗号的）读、8 句、…粗放进阶层式中，0-4 粗放进阶层式是层面型结构——涉及字内信息处理，属计算文字学研究范围，4-8 粗放进阶层式是线串型结构——涉及字间信息处理，属计算语言学研究范围。其中，0-6 粗放进阶层式属计算语汇学研究范围。以上是 **GTCM 的实例 1**。以下是 **GSCM 的实例 2**：在汉语中，由字的笔画构成的层面型结构按照 1、2、3、…、m 笔画数组成 1-m 精细进阶层式（在总量上等价于 GTCM 的 0-4 粗放进阶层式）；由字与字组构成的线串型结构按照 1、2、3、…、m 字（音节）数组成 1-m 精细进阶层式（在总量上等价于 GTCM 的 4-8 粗放进阶层式）。**GTCM 与 GSCM 一致的实例 3**：在英语的 0 字母、1 词头和词尾、2 词缀、3 词根、4 词、5（无虚词的）词组、6（有虚词的）短语、7 意群、8 句、…粗放与精细一体化进阶层式中由字母构成的线串型结构按照 1、2、3、…、m 字母数组成 1-m 粗放与精细一体化进阶层式。其中，**Z-ASCII** 为（完整）**子全域**。

（3）中文信息处理的**间接形式化**，即：（所有）**进阶层式**均**表格化**，**双列表**左列**数字化**、右列**字组化**。简称：三化。（为便于有针对性地计算或查询——如：处理数据、获取信息、重用知识，依据异义排列序趣简美法则及相对完全归纳原则和同义并列对应转换法则）设定 GSCM 与 GTCM 及 Z-ASCII 的（有限）目标域（如：汉语“字与字组细分”或英语“词与词组细分”）系列电子表格数字代码 m n 阵列。

（4）其中，**（有限）目标域** = （目标域内的所有用户的）**未知域** + **已知域**。已知与未知，相对于具体用户而言；系统（有限）目标域的设定，原则上涵盖具体用户的未知域与未知域。子全域是

用户和系统共同遵守的基准参照系——基准元素是相对完全的；超子域是系统定制的应对参照系——应对成员也是相对完全的。这是协同智能计算系统的**标准平台——**（有限）目标域——也是多**艾真体**设计的基准元素和应对成员的取材来源。用户及用户群的定制基准参照系和定制应对参照系，由其已知域及（用户可推测的）未知域构成的**（非常有限）目标域**，可在使用之前预定并在使用过程中通过人机交互而逐步优化并拓展[其上限是（有限）目标域]。

（5）（有限）目标域，由基准参照系（即：Z-ASCII）和应对参照系（即：GSCM 与 GTCM 及 Z-ASCII）构成，其中，**已知域**，是（有限）目标域中（用户）已知部分（涉及：用户或用户群的特征信息及其使用记录）；**未知域**，是（有限）目标域中（用户）未知部分。**GCM** 是 GSCM 与 GTCM 及 Z-ASCII 的总称。

3．2．测序定位——明确数据、信息、知识之间的数量关系，即： D = I + K

工程融智学研究证明：GCM 数据结构是确定的，已知或未知的超子域进阶层式成员一定是 GCM 的成员。只要（有限）目标域的数据确定，就能通过间接形式化的方式，计算或查询已知域的知识或搜寻未知域的信息。在间接形式化的前提条件下，数据、信息、知识之间的数量关系：

（有限）目标域（全部数据） = 未知域（未知部分的数据） + 已知域（已知部分的数据）

（有限）目标域（全部数据） = 未知域（信息或未知数据） + 已知域（知识或已知数据）

（数据 Data 的映射集） **D** = （信息 Information 的映射集） **I** + （知识 Knowledge 的映射集） **K**

公式中 **D** 表示 GCM 数据库 1-m 进阶层式双列表自动编号数字代码构成的 m n 数字阵列，**D** 是 **I** 与 **K** 两映射集之和，即：**D = m n = I + K** 。这是体现间接形式化的基本计算公式。

4．结果

公式（**D = m n** 和 **D = I + K**）反映了间接形式化约束条件抽象映射集之间的基本关系，其中，抽象的依据是 1-m 进阶层式双列表的左列数字和右列字组完全符合同义并列对应转换法则。也就是说，（有限）目标域 D（可计算的数字代码——指代与之同义并列语言文字），是计算机处理的对象；其中，未知域 I（将获取的信息）和已知域 K（可重用的知识），如：遵循相对完全归纳原则采集的（汉语的）字或字组或（英语的）词或词组，是自然人理解的对象。理解对象的形式[即： I 和 K（内容，钱币的一面）的间接记录 m n 与 D （形式，钱币的另一面）]与理解对象的内容（即：间接记录 D 与直接呈现 I 和 K）恰似钱币两面的关系。

4．1．自然语言处理的间接形式化——三化

自然语言处理的总量控制模型——形式化标准平台，由计算机数据库的一系列双列表表格的序列号 **m n 阵列**组成，其特征在于：**间接形式化**，（1）进阶层式**表格化——**记录自然语言文字的基础表格采用双列表，即：各进阶层式的多个双列**表序号 m** 异义排列，各双列表**行序号 n** 异义排列；（2）左列**数字化**、右列**字组化——**所有双列表的左列数字和右列字组（相对完全归纳，如：汉语的“层面型结构”和“线串型结构”；英语的“线串型结构”）之间逐行同义并列。简称：三化。

4．2．中文信息处理实施例

以中文信息处理为例，**“三化”**的特点是：**GTCM** 的 **0 进阶层式**（子全域平行层式）的一个双列表是基准参照系（含 ASCII 的 **Z-ASCII**），GTCM 的 **0-4 进阶层式**的五个双列表（成员是“层面型结构”）是字内信息处理的应对参照系（与 GB 或 Unicode 兼容），GTCM 的 **4-6 进阶层式**的三个双列表（成员是“线串型结构”）是字间信息处理的应对参照系（等价于 **GSCM** 的 1-m 个双列表）。

形式化标准平台，即：计算机辅助**选域定向**和**测序定位**的计算模型，由 m 个双列表构成自然语言处理的总量控制模型——数据库，其中，列表号 m 与行号 n 组成格号 m n 的数字代码阵列，非常便于自动化计算和查询；双列表的左列是数字代码、右列是图形符号；其特征在于：（1）间接形式化基于多个双列表的数字代码 m n 阵列，区别于所谓直接形式化的图形符号，双列表的数字代码行与图形符号行同义并列是转换的基础；（2）对汉语而言，图形符号的层解或串解形式，记录在进阶层式数据库的 m 个双列表的右列；（3）具体的层解信息和串解信息，通过建立一系列标注列与查询表而实现，并可同步建立用户查询记录表和索引表；（4）图形符号依据同义并列

对应转换法则在八大形式体系之间相互替代（如：汉语字符与汉语音节的相互替代）。

4．3．英文信息处理实施例

以英文信息处理为例，**“三化”**的特点是：**GTCM** 的 **0-m 进阶层式**与 **GSCM** 的 **1-m 个双列表**（成员是“线串型结构”）总量相等且形式一致。

4．4．子全域平行层式的间接形式化计算模型（工程化方法的基础）

子全域 Z-ASCII 是由 m 个双列表构成的平行层式数据库。其中，每一个双列表的图形符号列，由 n 个可枚举元素构成。GCM 的 0 进阶层式数据库中元素排列成 m n 阵列，选域定向：子全域平行层式的模式识别 1（形式消歧）——有直接与间接两种基本类型，前者比对双列表的符号图形列；后者比对双列表的数字代码列。测序定位：子全域平行层式的元素计量。

4．5．超子域进阶层式的间接形式化计算模型（工程化方法的基础）

超子域 GCM 是由 m 个双列表构成的进阶层式数据库。其中，每一个双列表的图形符号列，由 n 个可列举成员构成，目标域成员排列成 m n 阵列。选域定向：超子域进阶层式的模式识别[1（形式消歧）和 2（内容消歧——另文详解）]。测序定位：超子域进阶层式的成员计量。

5．结论

自然语言处理的总量控制模型——形式化标准平台的特征及优点：结构上，区分子全域的平行层式和超子域的进阶层式；**性质上，**先区分直接形式化与间接形式化——**化复杂为简单**，再区分算术级数（即：Z-ASCII 间接形式化）与几何级数（即： GSCM 和 GTCM 通过超子域进阶层式的一组双列表的数字代码“阵列 m n”间接形式化——**化几何增长为算数增长**）；**性能上，**区分**选域定向**与**测序定位**，其中，前者（即：区分平行层式和进阶层式的表号 m）把握方向——宏观处理，后者（即：区分平行层式元素和进阶层式成员所在的格号 m n）深入到位——微观处理。

在标准平台的底层 Z-ASCII（**GTCM 的 0 进阶层式**）和基础层（**GTCM 的 1-4 进阶层式**）优于 ASCII 和 GB 或 Unicode 中文信息处理。比较：在 GB 和 Unicode 汉语字符集中，因中文特有的字内信息（如：“层面型结构”信息）没有形式化，故无法计算或查询。在 GTCM 的 0，1-4 和 4-6 进阶层式中，因中文特有的信息（如：“层面型结构”与“线串型结构”及其关系的信息）**间接形式化**（有 GTCM 的 4-6 或 GSCM 的 1-m 和 GTCM 的 1-4 对 Z-ASCII——GTCM 的 0 的支持），故间接计算及直接呈现和查询都很方便。加之，兼容 ASCII 且改进并优化了 GB 和 Unicode 的汉语字符集，其产业化途径通畅。

在标准平台的底层、基础层和中层（**GTCM 的 0-6 进阶层式**），汉语的层面型结构与线串型结构的关系，即：中文特有的层解字内信息与串解字间信息，可在进阶层式数据库的多个双列表中得到完整记录或体现。具体的计算和重用，可通过设计多列查询界面而有针对性地实现。

在标准平台的上层和外围（**GTCM 的 7-12 进阶层式**），可通过记录和查询用户重用日志及其调用底层、基础层和中层的信息索引，实现计算机辅助研究。进而有针对性地调用记录和查询，获取相应的过程信息或应用信息。提取可重用的语言知识和领域知识以及常识也很方便。从而，为计算机辅助学习或进一步的研究，建立重用知识索引和获取信息索引及素材库。

Z-ASCII 与 ASCII 的关系：（1）后台切换的理想对接方式和前台切换的现实对接方式；（2）内外码统一；（3）软件切换和硬件切换。具体方式，须视具体需要而选用（技术细节省略）。

综上所述，我们认为：直接形式化有其特定领域或限制条件。在有限目标域确定[即：m n 阵列（与之对应的汉语“字与字组的细分”或英语“词与词组的细分”到位）数据明确]的情况下，（借助：标准平台）**间接形式化**不仅可使自然语言处理效率显著提高，而且，形式化难题也将迎刃而解。

参考文献

[1] 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 1997

[2] 邹晓辉：优化“语义信息处理”的新方法与实施例[A]CLSW-6[C]厦门大学 2005

[3] 陆俭明、郭锐：汉语语法研究面临的挑战[J]世界汉语教学 1998（4）

[4] 俞士汶：关于汉语信息处理的认识及其研究方略[J]语言文字应用（总第 42 期）2002（2）

[5] 邹晓辉：协同智能计算语言数据库的设计方法[J]潜科学（第 32 期）2004

[6] Zou Xiao Hui（邹晓辉）：THE GROSS CONTROL MODEL OF SEMANTIC VOCABULARY AS DICTIONARY WITH EXAMPLES [A] RECENT ADVANCEMENT IN CHINESE LEXICAL SEMANTICS [C](CLSW-5)Singapore 2004

[7] 邹晓辉：字与字组的关系——试论字本位理论的发展[J]潜科学（第 39 期）2005（1）

[8] 邹晓辉：默契通信与间接计算对自然语言处理的重要性[J]潜科学（第 42 期）2005（4）

[9] 邹晓辉：语义信息新论[J]潜科学（第 43 期）2005（5）

[10] 邹晓辉：一种知识信息数据处理方法及产品[J]发明专利公报 G06F163 知识产权出版社 2000（11）

[11] 陈肇雄主编：机器翻译研究进展[C] 1-564 页，电子工业出版社 1992

[12] 黄河燕主编：机器翻译研究进展[C] 1-282 页，电子工业出版社 2002

[13] 邹晓辉：字的形式化定义——试论字本位理论的根基[J]潜科学（第 28 期）2004（12）

[14] 邹晓辉：中文信息处理的新方法[J]潜科学（第 42 期）2005（4）

尾注

本文写作的过程中还有针对性地系统地参阅了以下网络文献：

AI in the news ©2000 - 2005

R.V.L.Hartley（哈特莱）.1928,Transmission of Information,BSTJ,Vol.7,p.535-536.

C.E.Shannon（申农）.1948, Mathematical Theory of Communication,BSTJ,Vol.27,p.379-423,632-656.

中国人工智能学会：中国人工智能进展（2003）[C]

WordNet，ILD，Longman Lexicon of Contemporary English，CYC，online version　树结构

信息科学交叉研究学术研讨会论文

两个基本信息公式及其算法的坏与好的比较

——指出：哈特莱-仙农提出的经典信息公式是坏算法

(强调：语义信息新论提出的基本信息公式是好算法)

摘要：为什么说哈特莱-仙农的形式信息公式是坏算法，而语义信息新论提出的基本信息公式是好算法？指出前者是坏算法与强调后者是好算法，有必要吗？本文首先从数学上回答了第一个问题，接着，从信息科学与计算机数据处理两个角度回答了第二个问题。仙农信息论深入人心，这是众所周知的，但是，哈特莱提出的经典信息概念及其基本公式是仙农信息论的基础，则往往是通信专业以外的人士所不清楚的。同理，哈特莱-仙农的形式信息公式都已普及，这也是学界认可的，但是，该基本公式是坏算法，却又往往是数学专业以外的人士所易忽视的。如果读者不理解这两个基本的知识要点，那么，也就必然认识不到语义信息新论提出基本信息公式的作用及其重要性。

关键词：哈特莱信息　仙农信息论　坏算法　好算法　语义信息公式

1.绪言

1.1.领域：本文探讨哈特莱-仙农提出的形式信息公式与语义信息新论提出的基本信息公式（简称：语义信息公式）的关系，即：$I=H-0=N\log S$（哈特莱“指数-对数”信息公式）及 $I=H_s(p_1,\ldots,p_n)-0=-K\sum p_i\log p_i$（仙农“对数-概率”信息公式）与 $I=D-0=m\,n$（邹晓辉“自然数-矩阵”信息公式，$I_U=I_D-I_K$ 中 $I_K=0$ 即不考虑知识及语义时的情形）的区别和联系（当 $H=I_D$ 时），属于信息科学的基本理论研究领域，涉及数学、通信和计算机科学的交叉研究。

1.2.特殊性：本文从语义信息公式的独特视角，深入透彻地分析信息（信息科学的核心概念的内涵及其本质和外延）及其计算模型（涉及基于双列表的间接形式化方法和序位恒等式及具体的算法优选）。

1.3.重要性：本文不仅关注信息论的基本问题——信息与信息量的关系，而且指出被忽略的两个基础问题：a 信息计量原理（语义信息公式和序位恒等式）；b 区别算法好坏的标准。

1.4.研究途经：首先，在战略层面，直接采用语义三棱模型和语义信息公式，明确信息的内涵及其本质和外延以确定定性分析的基础。接着，在策略层面，借助基于双列表的间接形式化方法，既可让八大形式体系各就各位，又可使自然人的定性分析擅长与计算机的定量分析特长各得其所且相得益彰。最后，在战术层面，明确知识信息数据的序位恒等式（即：$m_1 n_1 + m_2 n_2 = m n$）以确定定量分析的基础。其间比较了两种对付指数增长的基本思路——(由 Nyquist 提议，Hartley 采用 Shannon 沿用——我们至今仍在使用)对数与(邹晓辉采用)矩阵（结合关系数据库，不仅优选重用分布函数和线性代数方程及线性规划等现成的好算法很方便，而且还可直接使用自然数进行算术计算）。

1.5.局限性：一、哈特莱-仙农的形式信息公式的局限性；二、区别算法好坏的相对性。了解这两个限制因素，利于正确理解语义信息公式应用的具体限制条件。

1.6.基本假设：算法的好与坏以及算法的简单与复杂的全局判定比局部判定更重要。也就是说，我们说形式信息公式是坏算法而语义信息公式是好算法是全局判断。

1.7.贡献：进一步探讨了语义信息公式的内涵，如：a 明确提出知识信息数据的序位恒等式；b 明确提出信息计量公式（即：形式化描述好坏优劣）的科学评判标准：第一，可计算，第二，易计算，即：算法好，要么高效且足够简单，要么虽复杂但非常有效且经适当处理可化繁为简；c 明确指出基于双列表的间接形式化方法及其典型实施例的数学、通信和计算机科学的依据。

2.综述

我们知道，科学的信息概念及其数学理论渊源于现代通信技术实践。拥有共同的通信符号代码表（如：字母表、摩尔斯电码、ASCII、Unicode）是双方及各方通信的基础。至于各方如何具体编码或解码均可归结为：具体算法的选择。下面介绍前人（信息理论的先驱者）和我们先期提出的信息概念及信息量公式，同时，提出本文探讨的问题。

2.1.哈特莱（Hartley1928）提出的信息概念和信息量公式

概念 1：信息是（在通信符号表中）选择通信符号的方式。概念 2：选择的自由度[S 的 N 次方（其中，S 表示符号表中符号的个数，N 表示被选符号序列的长度）]可用来计算信息量的大小。公式 1：$I = N \log S$ 公式 2：$H = N \log S$ [I 或 H 均表示信息量 NlogS 是指数（即 S 的 N 次方）取对数的形式]。

分析与思考：问题 1_a“选择通信符号的方式”意味着什么？问题 1_b“选择的自由度”又意味着什么？问题 2_a“通信符号表”和“S 的 N 次方”意味着什么？问题 2_b 公式 1 和 2 是什么关系？分析 1：从语义三棱模型看，问题 1_a 涉及三个基本概念（范畴），即：“选择”（意）、“符号”（文）、“方式”（义）；一个复合概念，即：“选择方式”（意义）；一个复杂概念，即：“通信”，其中，“通”（物或载体载能的转换），“信”[即：（信息）=（意、文、义）]。问题 1_b 涉及：“自由度”[可选择的范围，即：（物、意、文、义）]。分析 2：问题 2a 涉及：“符号表”[双方或各方作出具体选择的共同依据或标准]，即：“符号”（文），“表”（义或体现具体关系的序位本义）。“S 的 N 次方”（表示信息总量呈指数增长，如：所有可能被选择的状态）。从语义信息公式看，问题 2_b 涉及：$I_D = H$（公式 1 和 2 的纽带）$I_K = 0$ （即：不考虑知识及语义）。结论 1：从全局上看，基于符号表的（人与人、人与机、机与机、机与人）通信皆受制于两个前提：1）算法的好与坏，如：算数增长与指数增长：2）算法的简单与复杂，如：整数与小数。结论 2：公式 1 和 2 的关系，即：$I_U = H - 0 = N \log S$ 。

2.2.申农（Shannon,1948）限定的信息概念和改进的信息量公式

概念 3：信息是用以消除随机不定性的东西。概念 4：信息量是随机不定性程度的减少。公式 3：$I_U = Hs（p_1，．．．，p_n）- 0$ ，公式 4：$H_s（p_1，．．．，p_n）= -K\sum p_i \log p_i$ 。

分析与思考：问题 3：“随机不定性”意味着什么？问题 4：“不定性程度”意味着什么？分析 3：问题 3 涉及“随机”（概率），“不定性”（歧义性）。分析 4：问题 4 涉及“不定性程度”（歧义程度）。结论 3：引入概率虽可使分析深入细化，但并不能也没有改变被取对数形式计算而掩盖的指数形式。结论 4：判定歧义性是定性分析；计算歧义程度是定量分析。请注意哈特莱-申

农的形式信息公式的区别与联系以及它们的局限性！沿该思路推广的后续者从根上也受其制约。

2.3.邹晓辉（ZouXiaoHui,1997）发展的信息概念、提炼的信息本质及语义信息公式

概念 5：信息的内涵，涉及四种类型：时空序位、质能序位、类例序位、数码序位。信息本质是序位本义（即：本真信息）。信息的外延，涉及三个论域或基本范畴：意（即：意识意向，如：知识）、文（即：符号形象，如：文本。物化的立体或活体为其特例）、义（即：序位本义，如：关系数据库中表格化的序位）。形式信息可由选域定位来识别。语义信息的性质判定涉及信息与知识的关系。概念 6：信息量可由测序定位来计算。形式信息的数量计算涉及信息与数据的关系。公式 5：语义信息公式：$I_U = I_D - I_K$（用于自然人或计算机用户的语义信息计量 1997-2005by XiaoHui Zou）公式 6：形式信息公式：$I_U = I_D - 0 = m\ n$（用于计算机的形式信息或数据计算 $I_D = m\ n$,1997-2005 by XiaoHui Zou）

继续思考：问题 5：“序位本义”是什么？问题 6：如何实现数据、信息、知识的统一计量？分析 5：问题 5 涉及（有限目标域的）“序位”。分析 6：问题 6 涉及（间接形式化的）“数据、信息、知识”。结论 5：区分有限目标域与任意目标域，可限定双方或多方通信或交流的论域，确保对话言之有物，交流文之有据，沟通思之有路，内心思之有理，且便于得到互联网及计算机辅助。 结论 6：以双列表的方式间接形式化的数据[如：八大形式（即：字、式、图、表、音、像、立体——静态虚拟、活体——动态虚拟）之一的具体形式（如：中文或英文）——有限目标域的数据，既可是未知域数据——信息，也可是已知域数据——知识]均可在未知域 I_U 和已知域 I_K 组成的目标域 I_D 的 m 列 n 行的矩阵 $m\ n$ 表格中选域、测序、定位。基于双列表的间接形式化与 $m\ n = I_U + I_K$ 相互相成。

2.4.问题汇总（提出进一步探讨的问题）

问题 7：为什么说哈特莱-仙农的形式信息公式是坏算法，而语义信息公式（注：公式 6 是公式 5 的特例）是好算法？即：公式 1、2、3、4、5、6 之间是什么关系？问题 8：指出哈特莱-仙农的形式信息公式是坏算法与强调语义公式是好算法，有必要吗？即：除指数取对数这一直接途径之外，还有其它间接途径可以获得更好的算法吗？

下面对问题 7 和问题 8 的分析（旨在寻找新途径）和解答（旨在开辟新途径）方式采用全新思路。

3.方法与结果

3.1.基本算法及思路的比较

首先从数学上解答问题 7，接着，从数学、通信与计算机数据处理的角度解答问题 8。

公式 1-4 和公式 5-6 分别表示信息量计算的两种不同思路。同样是计算信息量，公式 1-4 采用“指数-对数”及“对数-概率”的策略（无法回避直接计算很大的自然数乃至实数的问题），而公式 5-6 则采用“自然数-矩阵”的策略（可间接计算实数且可把很大的自然数分解为相当小之后再计算）。

当“$H = I_D$”且“$IK = 0$”时，“$I_U = H - 0$” 及“$I_U = Hs（p_1，．．．，p_n）- 0$”与“$I_U = I_D - 0$”等价且均为“$I_U = I_D - I_K$”的特例。公式 6 表示信息总量等于数据序位的总量，即：只考虑载体形式而不考虑承载内容（如：知识及语义），信息的总量等于信息熵的数量。

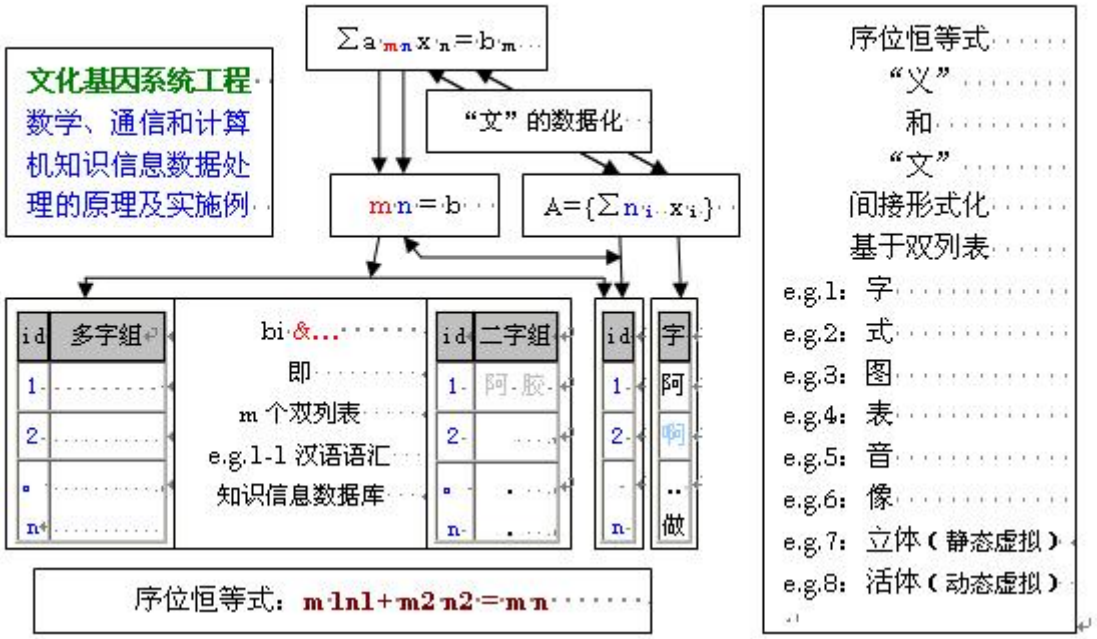
“$H = N\ log\ S$”及“$Hs（p1，．．．，pn）= -K\sum pilogpi$”与“$I_D = m\ n$”是对付指数增长（如：S 的 N 次方）的两种基本思路，涉及：(Nyquist and Hartley 采用)对数、(Shannon 引入)概率、(Zou Xiao Hui 采用)矩阵（优选重用分布函数和线性代数方程及线性规划等现成的好算法也很方便）三种算法类型，其中，前两种（“指数-对数”和“对数-概率”）属于“解超越方程的类型”与后一种（“自然数-矩阵”）属于“解线性方程（组）的类型”之间存在坏与好的区别。众所周知，对数与指数是函数与反函数的关系。对数虽可简化计算，却不能改变需要计算的数据总量呈指数增长的性质。因此，前两种基本算法都可归结为：基于指数的对数数值计算，不仅各次计算的数据总量庞大，而且，没有与知识信息数据处理对接的形式化途径。后一种算法则可归结为：基于自然数的算术数字计算，借助于双列表的间接形式化方法，不仅有与知识信息数据处理对接的形式化途径，而且，还是以量身订做的方式特制的，加之，事先已把呈指数增长的数据分解并转化成相应列或行的算术增长形式，因此各次计算的自然数数字的数量均相当限制，简单情况下仅需算术数字计算，复杂情况下还可直接重用各种现有算法且能好中选好、优中选优。由此可见，反映三种算法的两种基本思路相比较，后一

种思路在算法上的优越性是显而易见的。既然如此，为什么我们还要沿用前一种思路及其基本算法和具体算法的路径继续往下走呢？一则因为工作的连续性和程序的兼容性的需要；二则因为新旧两种基本思路的对接或转换也有一个磨合的过程。

现在看来，问题的关键，不仅是因为两种思路在选择数据（含：信息与知识）表示的形式化途径上不同，而且，还因为支持各自思路的信息观及方法论也有根本的区别。

3.2. 全新的思路和方法

“知识信息数据处理”方法的“选域、测序、定位”原理的“字、式、图、表”示意：



首先，从全局考虑：在战略层面，直接采用语义三棱模型和语义信息公式，明确信息概念的内涵及其本质和外延（见：本文 2.综述 2.3.概念 5）以确定定性分析的基础。

接着，在策略层面，借助基于双列表的间接形式化方法，既可让八大形式体系各就各位，又可使自然人的定性分析擅长与计算机的定量分析特长各得其所且相得益彰。

最后，在战术层面，明确知识信息数据的序位恒等式（即：由一个加群的表达式 $I_K + I_U = I_D$ 和三个乘群的表达式 $m_1 n_1 = K$ ， $m_2 n_2 = I_U$ ， $m\,n = I_D$ 相结合而构成一个环的表达式 $m_1 n_1 + m_2 n_2 = m\,n$）以确定定量分析的基础。

4.启示

4.1.数学思考

我们知道：算术增长与指数增长，在数学上是非常清晰的两种计算模式。从根本上说，算法好坏皆基于此。通常情况下，两者之间存在不可逾越的鸿沟。如仅限于数学思维，的确不易找到比对数再好的方法来对付指数增长。何况人们已习惯走“指数-对数”转化的老路。虽然也知道：矩阵是非常高效的数学工具，特别是在计算机辅助的情况下。问题是两者之间似乎没有连通的路径。因此，能否发现“基于双列表的间接形式化”这座化繁为简的“桥梁”（捷径）就成了“寻找好方法及好算法”的关键之所在。

4.2.信息科学思考（涉及：进一步的数学思考）——化繁为简的关键：信息计量原理

工程融智学所述“字（含：数字与文字及特殊字符）、式、图、表、音、像、立体（静态虚拟）、活体（动态虚拟）”八大形式体系的广义文本基因皆可依据子全域平行层式元素“异义排列序趣简美法则”及超子域进阶层式成员总量控制“相对完全归纳原则”和“同义并列对应转换法则”纳入“终极标准信息交换码”（*Z-ASCII*）文本基因“基准参照系”和 *GTCM*（文本总量控制模型）及 *GSCM*（音节总量控制模型）组合文本的“应对参照系”这一文化基因系统工程的总体框架之中。

4.2.1.子全域平行层式的例子——自然数的非常有限集

由个位（一位）数为元素而构成的二进制数的集合（仅有 *0* 和 *1* 两个元素）和十进制数的集合（仅有 *1,2,3,4,5,6,7,8,9* 和 *0* 十个元素）是子全域的两个极为特殊的平行层式——自然数的两个非常有限集。其共同的特点：一是元素个数非常有限；二是共享元素重用也仅限于子全域。

4.2.2.超子域进阶层式的例子——自然数的有限变换集

基于{*0,1*}与{*0,1,2,3,4,5,6,7,8,9*}的元素而组成的数字组合是超子域的多个进阶层式（其成员由一、两、…、多位数字构成）——自然数的有限变换集（限定在可计算且可接受的范围）。其共同特点：一是元素被重用的次数多、频率高；二是随着元素位数的增加会产生相应的进阶层式，如：由两位数、三位数、…、多位数构成的 *1*、*2*、...、*m* 进阶层式；三是每一进阶层式有位数相同的多个成员——具体的数字组合；四是子全域可视为 *0* 进阶层式。

4.2.3.典型例子——子全域（如：*S*）与超子域[如：*S* 的 *N* 次方（涉及：两种基本思路的比较）]

两种基本思路及其算法类型：简单转换后还需复杂变换的所谓直接计算（即：基于“指数-对数”的超越数计算，如：化“*S* 的 *N* 次方”为 *NlogS* 以对付“指数爆炸”）与复杂转换后只需简单变换的所谓间接计算 [即：基于“自然数-矩阵”的自然数计算，如：（限定在“自然数的有限变换集”）不仅“S 的 N 次方”而且“N”均可转换为“m 个进阶层式的 n 个成员”，从而，可更有效地对付“指数爆炸”]。

4.3.计算机数据处理——典型应用实例（中文信息处理的基本结构控制模型）

用 *m n* 表示进阶层式及其成员的数量：既可合一（如：*ASCII*）也可分别（如：二进制数的集合与十进制数的集合）建立序位恒定的子全域平行层式元素一览表（即：子全域 m0 的多个平行层式一览表，共有元素 n_0 的多个格）。如果由 *ASCII* 进一步发展到 *Z-ASCII*（如：增加汉字笔画这一特殊的平行层式），那么，由 $m_0\ n_0$ 矩阵可建立基于双列表的间接形式化元素符号模式自动识别的基准参照系（即：由子全域各个平行层式有限元素构成的数据集合，如：*Z-ASCII*）作为超子域进阶层式有限成员定量分析的基础。也就是说，通过生成、采集、比对、转换等方式，还可进一步建立基于 *Z-ASCII* 的序位恒定的超子域进阶层式成员一览表（即：*m* 个进阶层式一览表，共有成员 *m n* 的多个格）。如：*GTCM* 的 *0-6* 进阶层式成员的细化形式，由 *m n* 矩阵可建立基于双列表的间接形式化组合符号模式自动识别的应对参照系（即：由超子域各进阶层式有限成员构成的数据集合）作为选域（即：确定 m 的值）定位（即：确定 n 的值）的定量分析（即：测序——确定 m 和 n 的值）的数字计算依据。再进一步，从 *m n* 中选出已知域（即：$m_1\ n_1$）作为具体知识的序位集合，余留的未知域（即：$m_2\ n_2$）作为具体信息的序位集合，这样，基于“$m_1\ n_1 + m_2\ n_2 = m\ n$”的双列表间接形式化知识信息数据库可构成：标准化与个性化结合的一系列具体的通用和专用计算平台。*GTCM* 的 *0-6* 进阶层式是典型的应用实例。其中，字内信息处理限于 GTCM 的 0-4 进阶层式，字间信息处理限于 *GTCM* 的 *4-6* 进阶层式[（有关构造过程，由中文信息处理的间接形式化新方法具体介绍。见：参考文献）字外信息处理限于 ***GTCM*** 的 *7-12* 或 *7-15* 进阶层式（见参考文献，本文不讨论）]。具体计算模型：字内信息处理的文本结构控制模型[*STCM in word*（限于 *GTCM* 的 *0-4* 进阶层式）]：中文以笔画为基本单位建立 STCM（字内的层面型结构 1-m 细分进阶层式）。英文以字母为基本单位建立 *STCM*（词内的线串型结构 1-m 细分进阶层式）。字间信息处理的音节结构控制模型[*SSCM between words*（限于 *GTCM* 的 *4-6* 进阶层式）]：中文以字为基本单位建立 *SSCM*（字与字组的线串型结构 *1-m* 细分进阶层式）。英文以词为基本单位建立 *SSCM*（词和词组或短语的线串型结构 *1-m* 细分进阶层式）。

5.结语

5.1.结论：就知识信息数据处理的间接形式化方法而言，序位恒等式的发现，为语义信息新论从全局到局部优选好算法奠定了数学、通信和计算机科学的坚实基础。

“序位恒等式”和“信息基本公式”以及“间接形式化方法”结合，不仅在数学上的优越性可以发挥得淋漓尽致，而且，在知识信息数据处理上的优越性也可发挥得淋漓尽致。所以，工程融智学从数学、通信与计算机数据处理三结合的角度发现了：把几何增长形式分解之后转化为算术增长的新途径，即：把数据总量 I_D 分解为矩阵 *m n* 进而再把 *m* 列 *n* 行数字转化为自然数的算

术计算方法。这就造成了，全局算法上指数形式对数化的旧途径与几何增长形式算术化的新途径的区别与局部分析上信息熵的概率分析与信息量的分布函数分析的区别和联系。

5.2.总结：本文采用工程融智学的观点从数学、通信与计算机数据处理三个方面论述了形式信息公式与语义信息公式的关系。即：从数学上明确两组公式的本义（义），如：*S* 的 *N* 次方的数据 I_D 既可直接转化为 *NlogS* 也可间接转化为 *m n*——两种表示总量相等，算法各异。形式信息公式遵循的是指数增长的法则，对数计算是其简化途径。语义信息公式遵循的是算术增长的法则，矩阵方法是其简化途径。从通信上明确两组公式的用意（意），如：通信的符号表的数据结构不同，选择的效率当然也就不同。形式信息公式要求通信双方或各方从呈指数增长的一个数据表中做出选择。语义信息公式要求通信双方或各方从呈算术增长的多个数据表中做出选择。从计算机数据处理上明确两组公式的文本（文），如：数据结构不同，查询路径和计算效率也不同。形式信息公式要求文本或数据直接形式化，语义信息公式要求文本或数据间接形式化。

5.3.议论：坏算法特点是几何增长或指数增长（如：*S* 的 *N* 次方）；总量太大而失控（如：小数或有概率特征）；直接计量方法（由指数到对数的算法变换）；混合计算（如：混杂的字符集）。好算法特点是算数增长（如：*N*）；还可进一步分解为 *m* 个总量可控的进阶层式（如：整数或有周期特征）。间接计量方法（由指数到算数的间接变换）；分解计算（如：单一的字符集）。由有限目标域的一组双列表及其数字代码组成矩阵作为间接形式化方法的数学基础，其中，左列编号与右列数据一一对应。I_D， I_U ，I_K 在 *m n* 中的序位是一致的，即：{左列编号集合} = {右列数据集合}，其特征在于：有限目标域 D 由 m 列 n 行数字代码构成，即：$I_D = m\ n$ 或 $m\ n = I_D = I_U + I_K$。信息基本公式（定义式）与信息总量公式（计算式）的有机统一，不仅内容简明扼要，而且形式简捷高效。有了基准参照系与应对参照系的概念，数学公式、计算模型，在 *GTCM* 的 *0* 进阶与 *GTCM* 的 *0-6* 进阶之间，都可通过间接形式化的方式，借助计算机辅助和信息总量公式，实现间接计算和直接呈现。关键是认清信息的本质。自然人对算法的选择，影响计算机的具体处理方式。在不同进制数的集合 *N={...}* 中，除了以个位（一位）数为元素而构成的集合之外，还有以两位数、三位数、…、多位数为成员而构成的集合，这前后两组关系是值得进一步探讨的。同理，*ASCII* 或 *Z-ASCII* 与 *Unicode* 或 *GTCM* 的 *0-6* 进阶之间的关系，也值得进一步探讨。

参考文献

R.V.L.Hartley：Transmission of Information [J] BSTJ,1928 Vol.7

Claude E.Shannon and W.Weaver：Mathematical Theory of Communication [J] BSTJ,1948 Vol.27

N.J.A.Sloane and A.D.Wyner：Shannon,C.E Collected Papers［C］IEEE Computer Society Press1993

钟义信：信息科学原理[M]北京邮电大学出版社 1996

Losee, R. M.：A Discipline Independent Definition of Information［M］,Journal of the ASIS 1997,

邹晓辉：一种知识信息数据处理方法及产品[J]发明，知识产权出版社 2000

邹晓辉：协同智能计算语言数据库的设计方法[J]潜科学（第 32 期）2004（7）对北大、清华等介绍 2002

邹晓辉：协同智能计算知识数据库的设计方法[J]潜科学（第 39 期）2005（1）对中科院、清华介绍 2002

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

Zou Xiao Hui（邹晓辉）：The Gross Control Model of Semantic Vocabulary as Dictionary with Examples[A]Recent Advancement In Chinese Lexical Semantics [A] CLSW-5 [C] Singapore,2004

邹晓辉：重构“概念分类体系”的新思路与新方法（介绍“语义三棱模型”）[A] CLSW-6 [C] 厦门大学 2005

邹晓辉：优化“语义信息处理”的新方法与实施例（介绍“间接形式化方法”）[A] CLSW-6 [C] 厦门大学 2005

邹晓辉：中文信息处理的新方法（介绍“间接形式化”JSCL-2005)[J]潜科学（第 42 期）2005（4）

邹晓辉：“默契通信”与“间接计算”对“自然语言处理”的重要性[J]潜科学（第 42 期）2005（4）

邹晓辉：语义信息新论（介绍“信息基本公式”）[J]潜科学（第 43 期）2005（5）

理性的标准的协同智能模型

摘要： 有人（Martha Pollack）说：我们要构建智能的活动者而不只是智能的思想者。我们赞同这种观点。进一步考虑之后，我们继续向前推进。让所有的用户及其软件代理(艾真体)在文本总量控制模型或音节总量控制模型（即：标准平台）上工作或活动。意思是：用户能仅用母语理解程序、获取信息或重用知识，而艾真体则仅用 0 和 1 处理数据；双方或多方依据标准平台协同工作或活动。该标准平台基于终极标准信息交换码，就像过去和现在基于美国标准信息交换码或国际统一代码的情形一样。我们认为：让所有依据具体的理性智能模型设计的艾真体在一个依据统一的标准智能模型设计的通用平台上协同工作或活动，机器翻译的歧义难题将获得一个较为满意的系统化解决方案。

关键词： 软件艾真体　理性活动者　标准平台　智能模型

1．引言

在过去几年，人工智能的研究取得了长足的进展[1，“展望智能科学”（史忠植）[1]和“智能学：信息-知识-策略-行为的统一理论”（钟义信）[2]（http://caai.cn/documents/caai-10.exe）；Agents[3]（艾真体），Knowledge Discovery and Data Mining[4]（知识发现和数据挖掘）（http://www.aaai.org）；2，协同智能[5][6][7][8]和语义信息[9]（http://potentialscience.org）]。

然而也还有很多重要的问题没有得到满意的解决[3，不同的信息观之间的分歧依然较大，信息本质的理论探讨仍在进行（http://potentialscience.org）；4，知识表达和计量的问题仍然存在；5，智能的本质仍未搞清（http://caai.cn/documents/caai-10.exe）（http://www.aaai.org）AI©2000 - 2005]。

有鉴于此，本文提出一种协同智能的观点，试图通过“合理分工、开放互动、高度协作、优势互补的（基于融智学理论框架的）协同智能”在“强人工智能”与“弱人工智能”之间形成必要张力，并对“信息-知识-智能的理论”探讨中可能存在的问题提出一些值得深思的意见或建议，强调巩固根基。实质上也就是在人工智能与人类智能之间寻求一种和谐的解决方案。

概述：本文属于交叉-综合-公共-基础**领域**，具体涉及“信息-知识-智能的理论”探讨。其**特殊性**：直接采用融智学前沿的理论成果，探讨人工智能学界关心的“信息-知识-智能”问题。其**重要性**：从宏观上为解决强人工智能与弱人工智能之间的观念冲突提供科学理论上的疏导；从微观上为解决“消岐”[涉及：模式识别、语言理解、知识表达（典型实例：机器翻译）]的技术瓶颈提供科学理论上的支持（注：技术上的支持——标准平台，另文介绍）。**研究途径**：在回顾比较前人和他人的研究与自己前期的研究之间的异同的前提下，梳理融智学前期探讨的有关理论成果，旨在明确“信息-知识-智能的理论”的融智学探讨新思路：“理性人”的智能模式“由合到分”——突出艾真体的理性智能，“标准机”的智能模式“由分到合”——突出标准平台的标准智能，其发展就是让“协同网”的智能模式“融智整合”——突出计算机及其网络的协同智能。**局限性**：本文仅介绍新思路、新理论、新方法的基本框架的有关部分，具体细节和应用实例需阅读参考文献。**基本假设**：合理分工、开放互动、高度协作、优势互补的协同智能，可在强人工智能与弱人工智能之间形成必要的张力。**贡献**：在明确“信息-知识-智能的理论”探讨的新思路、新理论、新方法的基础之上，提出了：基于知识信息数据处理的融智学新范式的“消岐”新方案，即：理性的标准的协同智能模型，可让所有依据具体的理性智能模型设计的艾真体在一个依据统一的标准智能模型设计的通用平台上协同工作或活动，使机器翻译的歧义难题获得一个较为满意的系统解决方案。

2．“理性人”的智能模式

我们认为：迄今为止，探讨人工智能的三个流派（符号主义、连接主义、行为主义）和两种倾向（强人工智能与弱人工智能，或：“由上至下”与“由下至上”）之所以此消彼长且长期并存，有一个重要而深层的原因，即：实质上各自（这可能是学界和业界没注意到或注意得不够的方面）都在探讨“理性人”的智能模式或其某些方面。不仅探讨者会坚持“理性人”的某种立场（尽管时常会受情绪、脾气、怪僻等“非理性”因素的左右），而且探讨方式也在模仿“理性人”的某种智能模式。由于各个人的知识背景不同，加之各自处于认知发展的不同阶段，“理性人”对同一个问题不仅可能会得出合乎“理性”的答案，而且也可能会得出违背“理性”的答案。

2．1．什么是智能科学（或：智能理论）？

例 1（答案 1 陈述）：“展望智能科学”（史忠植）认为：“智能科学（作为）研究智能的基本理论和实现技术，是由脑科学、认知科学、人工智能等学科构成的交叉学科。”

例 2（答案 2 陈述）：“智能学：信息-知识-策略-行为的统一理论”（钟义信）（在阐述“信息－知识－智能的统一理论”时）写道：“智能理论作为 21 世纪最重要的两个学科——信息学与生物学——相互作用的交叉产物，是新世纪科学技术研究与发展的焦点。”

分析 1：关于智能科学（或：智能理论）的称谓与界定，例 1 和例 2 的问题：

问题 1：面对同样的学科领域，例 1 和例 2 给出的解释竟如此不同。不仅关于智能这门学科的称谓不同（这是次要的），而且涉及其知识来源的学科范围的界定也不同（这是主要的）。

问题 2：例 1 的作者是否知道“认知科学已包含认知心理学和人工智能这两部分”（是翻译失误，还是另有新的见解）？例 2 涉及其知识来源的学科（如：生物学）的范围是否界定过宽？

问题 3：（主要的区别在于）认知科学与信息学之间能视为等价或相同的学科领域吗？

即使排出关注焦点的差异，甚至忽略学者之间在智能观以及智能科学观的区别，也无法回避以上三个问题（指出它们有利于进一步的科学探讨！）。

结果 1：指出上述问题旨在提出新的观点，即：包含人类智能与人工智能的智能科学（或：智能理论），其知识来源至少涉及：脑与神经生理学、心理学（含：认知心理学）、（自然）语言学、（人工）符号学、逻辑学、数学、通信与计算机科学（含：计算机图形图像处理，语音处理，传感与遥测技术）等科学学科（甚至相关的哲学分支）。否则，就会犯思路过于狭隘的认知错误。

结论 1。知识背景的不同或个人知识的局限，是每一个学者，作为自然人，都无法回避的。面对一个复杂问题或复杂学科，自然人之间要达成共识，并不是一件容易的事情。

建议 1：采用协同智能的理念或策略，即：以“标准机”的通用智能模式，支持“理性人”的专用智能模式，协同辅助“自然人”进行科学探讨。如：确定知识门类划分、确定学科名称、确定各学科的核心概念和基本概念以及典型例证乃至相应的方法及工具等。

2．2．什么是智能？什么是信息？什么是“知识”？什么是“人工智能理论体系”？

例 3（答案系列陈述）：“智能学：信息-知识-策略-行为的统一理论”（钟义信）归纳概述，首先，把智能概略地定义为“认识问题和解决问题的能力”。接着，把智能理解为“一种有目的的行为”（本文称之为：“理性人”的“目的驱动论”）。于是，就得出“智能的完整定义”的以下系列表述：**定义 1　智能**，是在总体目的的驱动下，面对任何给定的环境，发现（定义）问题、确定目标、获得问题-环境-目标的信息、把信息提炼为知识、把知识激活为合理的策略、在策略引导下解决问题（满足约束）达到目标的能力。 **定义 2　人工智能**，是在给定问题、环境、目标的前提下，机器获取相关的信息、把信息提炼为知识、把知识激活为策略、并在策略引导下满足约束解决问题达到目标的能力。 **定义 3 本体论信息**，是关于“事物的运动状态及其变化方式”的直接表现，与观察主体的因素无关。 **定义 4 认识论信息**，是“主体所感知的事物运动状态及其变化方式，包括这种状态方式的形式（称为语法信息）、含义（称为语义信息）和价值（称为语用信息）”。语法信息、语义信息、语用信息三者的全体（我们认为：此处的“全体”应换为：“总称”），称为“全信息”。**定义 5　知识**是人们实践经验的结晶（最流行的知识定义）；经验，是有待确证的准知识。**定义 6　关于某类事物的“知识”**，是人们关于这类事物的运动状态及其变化规律的描述，包括这种状态和规律的形式（形态性知识）、含义（内容性知识）和价值（效用性知识）。**定义 7 常识**是被普遍公认因而无需证明的知识。**综合智能理论**，曾指出，知识激活成为（狭义）智能（体现为策略），需要有具体的求解问题、环境约束条件和问题求解的目标。否则就会成为空洞的智能。从知识激活的机制看，人工智能现存的三大学派（基于规则性知识的功能主义学派、基于经验性知识的结构主义学派以及基于常识性知识的行为主义学派）正好构成了有机互补的人工智能理论体系，不妨称为“广义人工智能”。它们共同的机制都包含“知识的激活”，只是由于知识的性质不同，激活的具体方法不同而已。正像其它（物质和能量）资源可转换一样，信息也是一类普遍存在的资源，可通过相应的加工机制

把它转换成为知识、策略和执行策略的行为，最终成为认知与行事的智能。信息是智能的源泉；智能是信息的归宿。这就是“信息-知识-策略-行为的转换与统一理论”。

2．3．本文对例 3 的探讨

2．3．1．分析 2：关于智能（含：人工智能）的定义，例 3 的问题

问题 4：众所周知，自然人一生中有相当多的情况是“非理性的”。试问：此时的自然人有没有智能？如果有，定义 1 就缩小了人类智能的范围。如果说“智能是一种有目的的行为；没有目的，谈不上有智能。”那么，“无目的随机应变能力”是不是一种“智能”？在“非理性”或“无意识”的情况下自然人有无“智能”？…值得进一步商榷。

问题 5：智能的本质是什么？仍然不清楚。定义 1-7 也没明确回答这个问题。

例 4：“智能的本质与定义[10]”（廉师友）认为“智能本质是信息对信息的一种恰当响应”，并给出以下定义“所谓智能，就是个体、群体或者系统能够对感知信息做出恰当响应的能力”。众所周知，由于信息的本质目前还存在争议，所以基于信息的智能定义及其本质也难下定论。

结果 2 及结论 2：我们认为：例 3 试图得到一般的“理性人”的智能模式。这与人工智能学界探讨的各种具体地模拟“理性人”的智能模式的思路方向一致但着眼点不同。众所周知，业界实际成功设计和应用的几乎都是后者——专业化的理性智能模式或具体类型。

2．3．2．分析 3：关于信息（定义 4 所谓“全信息”）的定义，例 3 的问题

问题 6：语法，既有形式方面，也有内容方面。例 3 定义 4 所谓“语法信息”只讲前者。

问题 7：语义，既有逻辑方面，也有语汇方面。例 3 定义 4 所谓“语义信息”只讲前者。

问题 8：语用，可指语言效用，更强调上下文。例 3 定义 4 所谓“语用信息”只讲前者。

结果 3：我们发现：例 3 定义 4 所谓“全信息”其实并不全。理由，上述问题 6-8 已指明其一；其二，为什么会这样呢？本文认为：其中一个主要原因，可能是“全信息”的提出者忽视或低估了自然语言的歧义性特征及其影响效力的深、广、久的特点。我们知道（但大众不知道或不认可）：例 3 所使用的“语法、语义、语用”三个词语是从符号学和语言学临时借用过来的，因此，全信息提及的语法、语义、语用不是大众所熟悉或偏好的通用含义。

议论：对此，例 3 的作者在撰写《信息科学原理》时是清楚的，但在与语言学和计算语言学乃至人工智能及计算机科学等领域的其他专家和大众（请注意：在面对“语法、语义、语用”这三个词语的时候，其他人完全可能选择其中两个意思的后者！！见：问题 6-8！）交流时却未必次次都能保证双方对此问题的认识是清楚一致的。这可能是造成双方对这个“全”的认知冲突的主要原因。不知这个原因的人（即使对作者本人而言，所谓“全”的假象一时也难以识破），可能会在欣赏“全”的同时却又说它（全信息理论）“遭遇到了实践上的灾难。”

结论 3：我们认为：理论的完整是相对的。如果一个理论的问题不能被发现，那么，它也就难以继续发展。“全信息”理论，没有认识到或至少忽略了这样一组事实，即：由“语法信息、语义信息和语用信息”中借用“语法、语义、语用”而必然造成的表达歧义——因为这三个词语本身就都存在概念上的歧义（见：问题 6-8！注：语法本质上是一种关系——词语之间的关系、语义至少有逻辑语义与词汇语义的区别、语用与上下文有关而具体使用时才与具体的人有关）。

2．3．3．希望首先要坚固“信息-知识-策略-行为的转换与统一理论”的根基

分析 4：由于前述“智能理论、智能、信息的定义”的问题，例 3 关于知识的定义和关于人工智能理论体系的整合，在理论的根基上也就必然存在相应的问题。

结果 4 及结论 4：我们不仅认同“信息-知识-策略-行为的转换与统一理论”追求卓越和完美的探索精神，而且，高度评价“（狭义）智能（体现为策略）”的观点，同时，也十分欣赏其试图融合人工智能三大学派的努力。不过，还希望首先要坚固其理论的根基！

2．4．启示 1

以上仅仅是对例 1-3 或两位学者的研究提出了 1-8 个值得进一步探讨的问题。类似的疏忽或失误，每位（包含我们自己在内）自然人学者都随时可能遭遇。正因为如此，我们更加坚信：通过

“合理分工、开放互动、高度协作、优势互补的协同智能（狭义部分）的方式”必然会尽可能地降低自然人学者个人的疏忽或失误及其给学术界可能带来的损失或负面影响。

针对以上“理性人”的智能模式，以下提出：“标准机”的智能模式。旨在“取长补短”。

3．“标准机”的智能模式

3．1．深入探讨

“理性人”的智能模式，是基于自然人的智能模式。鉴于自然人智能观的多样性，计算机中体现“理性人”的智能代理——“艾真体”（Agents）的模式也必然是多种多样的。

如果基于自然人的“理性人”的智能模式（由合到分与由分到合孰优孰劣的问题）还有待进一步的科学探讨，那么，能否先设计一种基于计算机的“标准机”的智能模式（由合到分与由分到合不仅孰优孰劣一比即知而且完全可以协同互补）来辅助自然人继续探讨“理性人”的智能模式，并实验构造“合理分工、开放互动、高度协作、优势互补的协同智能（狭义部分）”呢？

（我们设计的）文本总量控制模型（GTCM）与音节总量控制模型（GSCM）就是基于计算机及其互联网的“标准机”的通用智能模式。经过近几年（2000-2005）小规模试验和局部试用，其效果非常显著。事实证明：无论是知识，还是信息，只要在基于 GTCM 与 GSCM 的数据库和数据仓库中，自动查找，自动翻译，辅助训练、写作或创作，…，都是十分方便而高效的。

3．2．分析比较

缺乏 GTCM 与 GSCM 系统全面、明确持久、准确有力的支持，无论是自然人（含：个人和群体）还是计算机（含：单机和联网），其内（如：计算机的艾真体和自然人的神经细胞及组织）外（含：计算机的互联网和自然人的社会关系网）环境之间必然隔阂重重（这就是各种自然语言和人工语言以及程序语言应运而生的条件），人与机、机与机、机与人、（即使在网络和计算机辅助条件下）人与人之间也难以实现默契通信（众所周知，就交流或沟通而言，是否默契差别很大。谁都愿和默契的伙伴相处或在默契的团队里学习、工作、生活）。反之，则可实现默契通信或自动消歧。

体现理性智能的专用模式（如：艾真体）虽不错但难以协同运行，如有标准智能的通用模式（如：GSCM 与 GTCM 及 Z-ASCII）作为其底层支撑体系，则各种各样的艾真体乃至其用户也就有了共同的基准参照系和应对参照系。这样，可随时随地共享知识信息数据处理平台的全面支持或辅助，从而，也就自然避免了信息孤岛及其遭遇的歧义困境。

3．3．几组问题（涉及智能本质探讨，仅供思考，希望对读者有所启示！）

第一组问题：人工智能与人类智能之间究竟是一种什么样的关系?人工智能的强与弱之间又是一种什么样的关系?“协同智能”概念的必要性和重要性，何在？

第二组问题：简单性与复杂性是认识智能现象及其本质时常遭遇的一对矛盾。简单化与复杂化是设计智能系统时常遭遇的一对矛盾。简单与复杂是智能化进程中必须解决的一对矛盾。简单与复杂的关系是智能本质探讨的过程中必须处理的一个重要问题。

第三组问题：“人与机、机与机、机与人、人与人”之间的交流或沟通，为什么总是隔阂重重？协同智能计算模型 = 标准计算模型 + 理性选择模型，有何独特功用？

3．4．解题方法（“协同智能”和“协同智能计算模型”及其“标准计算模型”）

3．4．1．理论方法

首先，通过“计算机的艾真体”和“计算机的互联网”容易证明电脑智能是协同智能的一类特例；接着，通过“自然人的神经细胞及组织”和“自然人的社会关系网及组织”容易证明人脑智能是协同智能的另一类特例；最后，指出：虽然其它“人工装置”与“自然实体”也都可视为协同智能的各类特例，但本文仅限于介绍：由人脑智能与电脑智能这两类特例构成的狭义的协同智能，其特征在于：合理分工、开放互动、高度协作、优势互补地融智。

3．4．2．工程方法

首先，构建“合理分工、开放互动、高度协作、优势互补的协同智能”的总量控制模型（系统设计已完成且试用效果好），即：文本总量控制模型（GTCM）与音节总量控制模型（GSCM）。

其次，确定“理性人群”与“标准机群”（即;“理性人” + “标准机”）之间默契通信的间接形式化（数学化且计算机化）标准格式[（即：$I_D = m\ n$ 全域数码），其中 m 和 n 均为自然数，在基于计算机的“标准机”通用智能模型的数据库及数据仓库中 m 是两列表的序号而 n 是记录行的自动编号]。

再次，计算机系统的“（有限）目标域”数据（D）按标准格式实施全域数码化改造。

接着，一方面，自动构造“标准机”的通用智能模型；另一方面，协同构造“理性人”的专用智能模型。对信息获取与知识表达而言，前者可自动生成；后者可协同采集。生成、采集、比对、转换，有时（如：不确定时）可协同进行，有时（如：确定时）可自动进行。

最后，用户在“（有限）目标域”中重用“已知域的知识”或发现“未知域的信息”。

4．结语

什么是协同智能？协同智能，旨在“人类智能”与“人工智能”（或“强人工智能”与“弱人工智能”两种“冲突的智能观”）之间寻求一种张力，进而发展出一种“和谐的智能观”。理论上指：侠义融智学定义的顶级的智能观，如：融通、融合的智能观，即：基于“人与机、机与机、机与人、（借助计算机及互联网的）人与人”的“合理分工、开放互动、高度协作、优势互补”的“协同智能”观。工程上指：“协同智能计算系统”。

分析 5：由于看待“人工智能与人类智能的关系”这一基本问题的立场不同，造成了“冲突的智能观”（强人工智能、弱人工智能）与“和谐的智能观”（协同智能）的根本差异，其中，强与弱的人工智能观基于对立或争斗的立场；协同智能观基于融通或融合的立场。

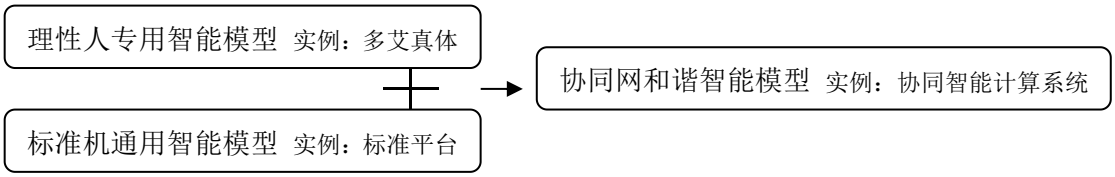
结果 5：见：三组关系，即：

协同智能 = 理性人的智能（专业化分工） + 标准机的智能（规范化协作）

协同智能计算模型 = 理性人的专用智能模型 + 标准机的通用智能模型

协同智能计算系统 = 理性人的专用智能代理 + 标准机的通用智能平台

以下框图是本文概括的三种智能模型及其典型实例之间的相互关系示意图。



结论 5：断言：人脑智能与电脑智能是协同智能的两类特例。即：由人脑智能与电脑智能这两类特例之间的特殊关系构成的协同智能与前两者的关系，是：进化发展的关系。

结果 6：发现协同智能与其两类特例（人脑智能与电脑智能）之间实现默契通信的条件——建立“标准机的通用智能平台（即：GSCM 与 GTCM 及 Z-ASCII）”获得 CA 全方位全过程的支持。

结论 6：预言 1：协同智能的时代初级阶段（狭义的协同智能得以认同的时代）即将到来，而高级阶段（广义的协同智能得以认同的时代）则需在协同智能计算系统普及之后才有可能到来。预言 2：继形式信息革命（即：数字化或数据革命。它是前不久和当前计算机和互联网时代的典型特征）之后的语义信息革命（即：由数据革命进入数字化知识革命。它是当前和近期未来的基于计算机和互联网的数字化知识经济时代的典型特征）将成为进入协同智能时代初级阶段的重要标志。

推论：智力的技艺体现，智慧的哲学体现，智能的科技体现，前述（智力、智慧、智能）三智融合的融智学体现，是人类认识“智”这一现象的四个发展阶段。

参考文献

1．史忠植：展望智能科学[A]中国人工智能进展[C]北京邮电大学出版社 2003

2．钟义信：智能学：信息-知识-策略-行为的统一理论[A]中国人工智能进展[C] 北邮出版社 2003

3．Software Agents[C]ISBN 0-262-52234-9 AAAI Press

4．Advances in Knowledge Discovery and Data Mining[C]ISBN 0-262-56097-6 AAAI Press

5．邹晓辉：一种知识信息数据处理方法及产品 [J]发明，知识产权出版社 2000

6． Zou Xiao Hui （邹晓辉）： The Gross Control Model of Semantic Vocabulary as Dictionary with Examples[A]Recent Advancement In Chinese Lexical Semantics [C](CLSW-5)Singapore,2004

7．邹晓辉：重构“概念分类体系”的新思路与新方法[A]CLSW-6[C]厦门大学 2005

8．邹晓辉：优化“语义信息处理”的新方法与实施例[A]CLSW-6[C]厦门大学 2005

9、邹晓辉：语义信息新论[J]潜科学（第 43 期）2005（5）

10.廉师友：智能的本质与定义[A]中国人工智能进展[C]北京邮电大学出版社 2003

尾注

本文写作的过程中还参考了以下网络文献：

AI in the news ©2000 - 2005

中国人工智能学会第 10 届全国学术年会论文集

钟义信：信息－知识－智能的统一理论（ ）

陆汝钤：发展知识工程 建立知识产业( )

邹晓辉：广义文本[现象（形式和内容）]与序位本义[本真信息（本质）] [J]潜科学（第 43 期）2005（5）

邹晓辉：融智学应用实例[J]潜科学（第 43 期）2005（5）

潜科学 2002-2005 各期有关信息、知识、智能的探讨文章（ ）

信息学基础研究

【摘 要】信息学基础研究在定性和定量分析的前提下集中论述（信息的）内容（意）、形式（文）、本质（义）等问题，并在信息形式化、部门信息学以及一般信息学等领域提出“语义、信息与智”的统一理论（框架）. 文章主要探讨信息概念（定义和分类）并采用协同智能科学的观点和方法，对当前有代表性的几个较为典型的信息观和方法论作了点评。

【关键词】一般信息学，协同智能科学，语义三棱，信息方程，基础研究[i]

一、引言

最近几年，信息形式化和部门信息学诸学科，取得了很大进展。首先，基于“奈魁斯特、哈特莱及申农的通信信息理论、维纳的控制论和图灵及冯诺依曼的数字计算机理论”而发展的信息形式化数字技术，已为通信、自动控制和计算机工程实践所证明，非常有效。接着，数字技术的广泛应用，大大促进了部门信息学相关学科的形成和发展，其景象如雨后春笋，其结果是信息学科化，其影响十分广泛。在形式信息革命大潮下，学界活跃分子不禁纷纷反思：信息为何如此神奇？日常生活通用的信息概念与各个科学领域专用的信息概念能统一吗？一个人，除了熟悉本领域专用的特殊信息概念之外，对其他领域的特殊信息概念和公共领域的一般信息概念是否也都可界定清楚呢？于是，哲学领域信息学转向呼之欲出，科学领域统一信息理论以及信息科学交叉研究领域一般信息学也应运而生（酝酿中）。

与部门信息学的发展相比，一般信息学的进展却相当迟缓，很多重要的问题至今没有得到满意的解决。如：信息哲学提出：信息、语义、智（含：智慧、智力、智能）等概念需要系统阐述，统一信息理论提出：部门信息学与一般信息学的关系需要清晰界定，信息科学交叉研究提出：一般信息的概念、原理及方法也需要系统研究。处于一般信息学前沿，既要明确常识到部门信息学和相关技术领域的信息概念，还要探究哲学、科学学乃至艺术等领域的信息概念。

有鉴于此，本文提出一种“语义、信息与智”的统一理论（框架），试图探究信息的内容、形式、本质等基础和核心的理论问题。同时，对当前信息科学交叉研究领域普遍存在几个认识误区，提出简明扼要而又富有启迪或值得深思的意见或建议。希望这一信息学基础研究成果对一般信息学理论探索者们有抛砖引玉之功效！欢迎同行多提宝贵意见！不当之处敬请指正！

二、正文

1.信息学基础研究立足于协同智能的科学信息观、方法论和相应的信息处理原理及方法。

我认为：无论是人类智力，还是人工智能，其实就是指信息处理能力。继人类与人工智能之后出现的协同智能在互联网及计算机辅助的自然人和软件工程支持的计算机之间互助互补的基础上获得了空前的发展。知识信息处理方法及产品广泛普及的现象，就是通过人类的自然之智与人工之智“合理分工、优势互补，高度协作、优化互动”的融智原理、方法及实例而得以体现出来的，具体表现为人机交互作用（Human Computer Interaction）。

融智信息观（集中探讨“信息的内容、形式、本质”）是基于协同智能而表达的“语义、信息（含：知识、语义信息、数据）与智”的统一理论（框架）。

融智方法论（集中探讨“间接形式化和一体化管理”）是发现并补足了在还原论和整体论之间长期缺乏中间环节——域位论之后而发展起来的协同智能方法体系（统一的理论框架）。

融智信息处理原理及方法，就是协同智能系统的逻辑推理和数学计算的原理以及知识信息数据处理方法。互联网及计算机辅助和软件工程以及人机交互作用可视为其发展的初级类型。

我还认为：信息学具有理论与实践紧密结合的特点。因此，我所进行的信息学基础研究，始终是沿着应用基础、工程基础、理论基础（简称：三基）三部曲进行的。因此我的研究成果，首先由一种知识信息信息处理方法及产品的发明（1997-2000）而获得实质性进展，进而提炼出理论。正如科学革命史一书的作者科恩所总结和预言的那样，我亲身体验“思想革命、口头革命、纸面革命”的过程。经过 2000-2005 六年尝试，我感受到“科学革命”不是一件容易的事情。尤其是采用知识信息数据处理方法和信息学理论融“三基”于一炉的系统工程，其困难可想而知。更不用说担当知识信息数据处理系统工程总体方案或蓝图设计者这样的艰巨任务。于是 2006 年我决定把工作重心收敛到“纸面革命”的过程控制上并突出人机交互过程的基础控制。因为，人机合作的观点及方法可对（静态的）语义、（动态的）信息及其处理之智进行深入仔细和富有实效的科学探讨。

“语义信息新论”一文介绍了“数据、语义信息、知识”的关系式和语义信息的定义式。“两个基本信息公式及其算法的坏与好的比较”一文介绍了哈特莱及申农的形式信息量的直接计算公式和邹晓辉的形式及内容信息量的间接计算公式的区别及联系。“广义文本与序位本义（本真信息）”一文介绍了语义三棱模型（“重构‘概念分类体系’的新思路与新方法”一文介绍了由语义三角到语义三棱的学术渊源，“字本位与中文信息处理”一文介绍了基于双列表的分层集合与基于多列表的标志集合以及新的方法论基础——域位论）。以下仅做有关概述。

2.“语义、信息与智”的统一理论（框架）

2.1 领域：“语义、信息与智”的统一理论（框架）属于信息学基础研究领域。

2.2 特殊性：信息科学原理作者钟义信把信息比作多面体。这更凸显了一般信息学的困难。我的设想是这样，即：部门信息学诸学科好比从不同角度直接观察与分析这个多面体，（因为有协同智能的概念）我的策略及方法是：重点研究（语义）三棱锥、（信息）四面体、（智的）四要点，进而可借助协同智能系统高效率地间接观察与分析（语义）多棱锥、（信息）多面体、（智的）多要点。这样，自然可做到以简驾繁进而有望开辟一般信息学研究的新途径。

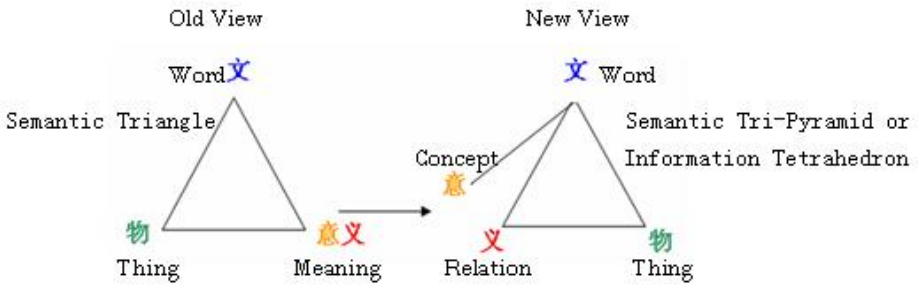
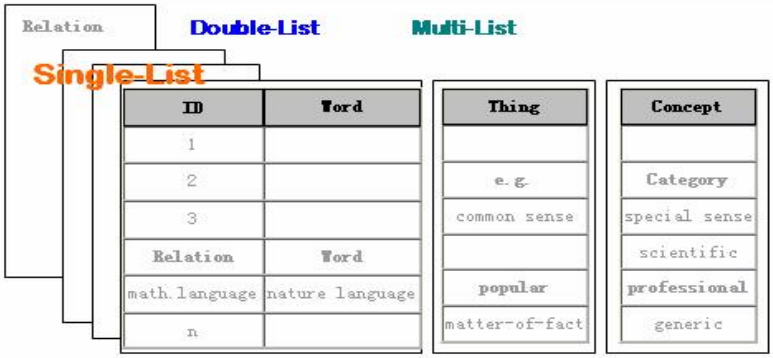


图 1 是从“语义三角”到 “语义三棱”的(几何模型) 示意图。英文仅表示模型的实例。

2.3 重要性：众所周知，任何一个复杂的多面体都可归结为若干个简单的四面体，四面体是最简单最基本的多面体。毋庸置疑，掌握（信息）四面体可为更有效地探知（信息）多面体提供理论上的基础性指导。这比盲人摸象式地直接研究（信息）多面体的常规做法更加可取。理由之一：研究（信息）四面体可得唯一而确定的结果及结论，而研究（信息）多面体可得出多种多样不确定的结果及结论。与其直接为不可为之事，不如先为可为之事（即：打好基础、创造条件），进而，再把不可为之事化为可为之事而为之。如：先熟悉“三位一体”简单变换，然后，再借助计算机去探究“多位一体”的复杂变换，自然容易的多。一旦众人掌握“语义、信息与智”的统一理论（框架），对深入研究“信息、语义与智”的细节，对理顺部门信息学各个学科与一般信息学的关系，对研究一般信息学的框架和细节，就有了高屋建瓴的行动指南。这时，再借助互联网及计算机辅助知识信息数据处理方法及工具的支持（如：间接形式化方法及其系列产品乃至一体化管理方法及其系列服务），就可为整个信息学体系全方位全过程探讨创造更有利的条件。那样，标准化与个性化兼容的信息概念体系的总论及各论，也就可望早日建立建全。至少可加速一般信息学同仁达成共识（如明确研究对象、方法及任务）的进程。

图 2 是“间接形式化”和 “一体化管理”方法及其工具的(表格模型) 示意图。



2.4 研究途径：（语义）三棱锥、（信息）四面体、（智的）四要点“三位一体”的几何模型（见：图 1），可形象地概括“语义、信息（知识、语义信息、数据）与智[信息处理（分与合）机制]”的统一理论（框架）；“信息方程”的代数模型，可抽象地概括“基于双列表的分层集合以及关系数据库的软件工程”的知识信息数据处理（方法）的序位模型（见：图 2）。

方法论和基本方法，涉及（相对完全）归纳、（完全）演绎、（间接）计算。下面以中文信息处理和知识工程为例，说明“知识、语义信息、数据”的关系及其处理的基本方法。首先，通过对（自然人与计算机）通用的标准文本（如：国际统一编码 Unicode）的定性和定量分解，提炼出单一集合的各个子全域（如：字母表、笔画表、数字表、各种特殊符号表），进而，区分出分层集合的各个超子域的成员所归属的各个进阶层式（如：中文的“一，二，…，多”笔画的字——层面型结构和汉语的“一，二，…，多”音节的字与字组——线串型结构），以此作为对“语义信息和知识”实现“间接形式化”的基础（见：图 3），然后，按照“标准化与个性化兼容”的原则，对进一步提炼出基于各个学科分类的标志集合（如“语言文字、通用常识、专用知识”分科标注），实施“产、学、研、用、算”一体化管理，以此可实现“语义信息和知识”的获取、表达、以及有针对性地重用。此方法的原理和实施例，在“字本位与中文信息处理的基础”专著中将详细介绍。这里只概要介绍其中与本文有关的方法论和基本方法。

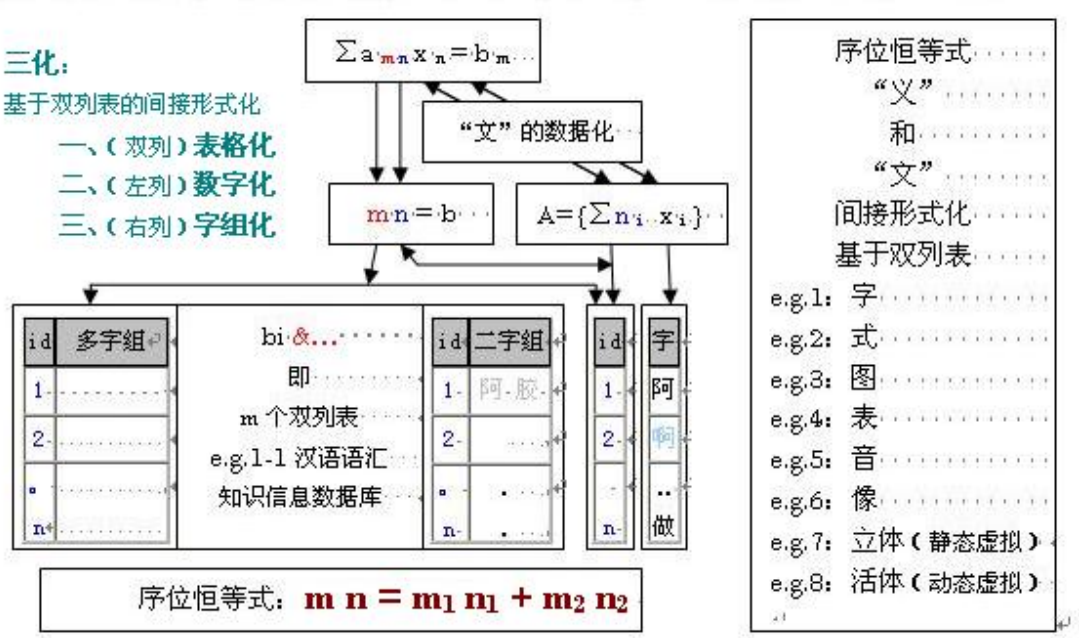


图 3 是以字与字组的关系为实例表示的“间接形式化”（三化）示意图。

“广义文本”（字、式、图、表、音、像、立体、活体）由“字”这一“狭义文本”推广而来。其中，涉及“文”与“物”两个基本范畴是显而易见的，而“意”这个范畴作为“主体的选择”必然可由“文”与“物”所蕴含或表达，这不证自明，可理解的难点在于“义”作为控制“物、意、文”的根本范畴（本真信息）似乎只能以逻辑和数学的方式予以明确或掌控。“意义=意+义”的内涵及问题，至今仍未被人们所普遍重视。于是，由“广义文本”（物、意、文）再聚焦到“本真信息”（义）的过程，并非任何人（尤其在缺乏相应背景知识和创造能力的时候）所能快速驾驭并收敛到位的。好在协同智能的出现，带来了本真信息充分共享的希望。

日常生活中的“集合”可这样被简化成为数学上的“集”。

方法论：	整体论	各就各位论		还原论
集合分类：	杂多集合	标志集合	分层集合	单一集合
		商集	直集	集
元素：	任意的复杂对象	抽象的标志的值	逐级分列	简单对象个体的数
		定性分析	结构分析	定量分析
实例：	多语词典	分类词典	定位词典	字母表

图 4 是“方法论”与“集合分类”（相互关系）示意图。各就各位论是域位论的局部用例。

图 3 和图 5 的模型呈现一个基于 GLPS（全球语言定位系统）的 GKPS（全球知识定位系统）共享共建全域平台，以“产、学、研、用、算”人机协作为特征的融智过程（涉及文化基因工程）是标准化与个性化结合的开放式进化发展平台，其作用：既方便计算机辅助教学；又方便广大师生与遍布各级学校乃至校外的网络化计算机系统之间的大协作、大融通、大融合（涉及一种生产式合作型教学及科研的模式，有利于改掉各级师生乃至各类各级专家常有的孤芳自赏、固步自封或闭门造车的陋习）。

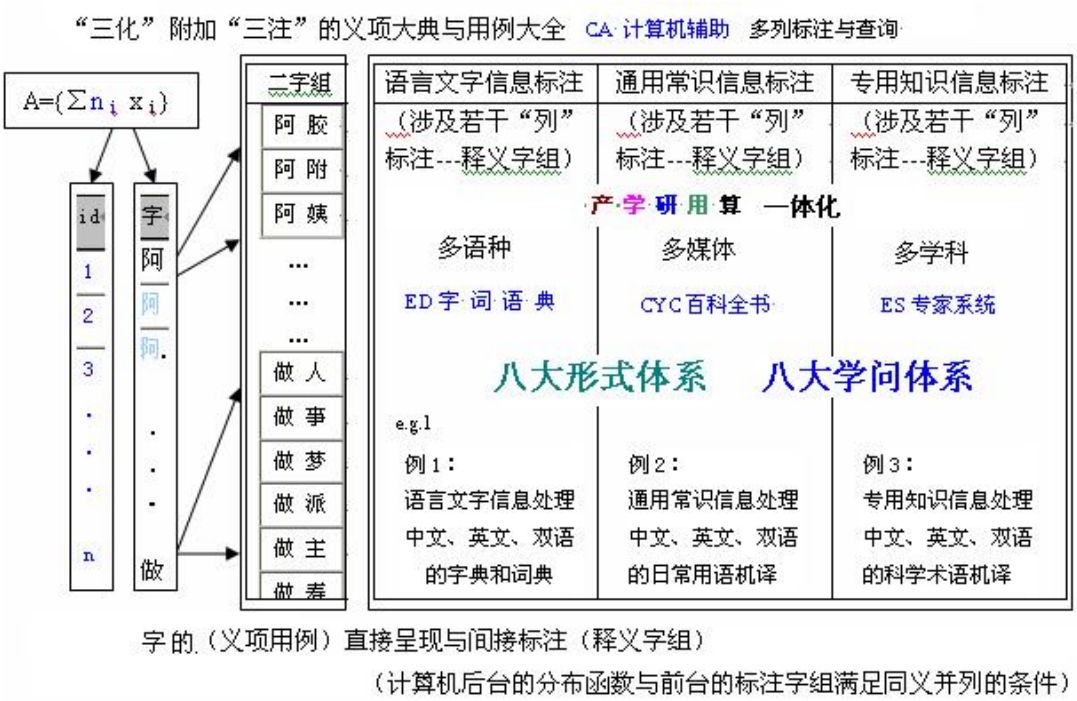


图 5 是以字与字组的关系为实例表示的“一体化管理”（三注）示意图。

2.5 基本假设（局限性）：由于对信息多面体的归纳、演绎、枚举、类比乃至计算、统计、估计必受人的认识认知与实践优化的进程限制，因此，由信息多面体到信息四面体的收敛及其逆过程（发散）建立在间接形式化和相对完全归纳的逻辑推理及数学演算的基础之上。本研究对信息的讨论及处理，设订了可计算且有限的目标域，即：目标域=已知域+未知域，$I_D = I_K + I_U$ ；目标域=多表*多格，$I_D = m * n$ 或 $I_D = n * n$（间接计量单位：ge）。

图 5 是本真信息（义）、形式信息（文）、内容信息（意）的（逻辑与数学）关系示意图。

2.6 可能贡献的创新知识点：

2.6.1 信息概念体系有三个层次，一，信息的现象和本质（序位）；二，现象的形式信息（即：数据，计量单位 bit 比特）和内容信息；三，内容的已知部分（即：知识）和未知部分（即：语义信息）。间接计量单位 ge 格。如：当每个格仅占 1 比特时，bit 就是 ge 的特例；当每个格仅有一个知识点时，个就是格的特例（即：形式信息与内容信息的非对称可忽略不计）。

2.6.2 信息学基础研究把一般与特殊兼容的信息定义在形式上可数字化，内容上可知识化，本质上可序位化的未知域，其中，“三可”属于理论方面的基础研究领域；“三化”属于实践方面的基础研究领域；基本分类有：形式信息（文）、内容信息（意）、本真信息（义）。

2.6.3 广义文本，体现：信息形式的丰富性（字、式、图、表、音、像、立体、活体），其中，既涉及：物（质、能、时、空）的“虚拟映射”并子集（物象信息），又涉及：意（含：静态的“知”和动态的“情、意”及其交融的“个性”化“选择”）的“虚拟映射”并子集（含：潜在的意向信息和显现的意识信息）。

2.6.4 一些值得深入探讨的新课题：

1）语义信息和知识是可通过数据而间接计量的

$I=H=N\log S$（Hartley）基于指对数转换的熵和形式信息计量公式，$Hs(p_1,\ldots,p_n)=-K\sum p_i\log p_i$（Shannon）基于概率的熵和形式信息计量公式，与 $I_U=I_D-I_K$且 $I_D=n*n$（ZXH）基于双列表分层集合及关系数据库的语义信息计量公式相比较，在假设 $I_K=0$的同等条件下，后者间接形式化的计算结果等价于前两者且有可直接使用一切已知算法的优点。

2）信息量= 能量与质量的比值

（表示 n 个表和 $n*n$ 个格的）序位模型 $n*n$ 在形式上与光速的平方 $c*c$ 相似，提示：当 $I_K=0$ 且 $n*n=c*c$ 时，可由 $I_U=I_D-0=n*n$ 和 $E/m=c*c$ 推知“信息（量）、能（量）、质（量）”的关系式 $I_D=E/m$ 即：能（量）与质（量）的比值就是物的自然信息（量）。

3）信息方程

狭义信息方程 $I_U=I_D-I_K$（即：语义信息定义式）是（间接形式化条件下）“语义信息、数据、知识”关系式的变形。它是“获取（语义）信息、处理数据、重用知识”实施总量控制的理论基础。在科学理论上揭示（语义）信息量、数据量、知识量三大变量之间的相互关系，可得：信息学基础研究领域的三大基本原理。在技术实践上揭示知识信息数据处理方法，可得：数据（间接计算与直接呈现）及知识（间接计量与间接呈现）处理的高效方法。

（满足条件 $w=a+bi+cj+dk$ 的）广义信息方程 $f(x,y,z,ict)=w$ 似乎可揭示“空间、时间、能量、质量、信息量”五个变量的相互关系，而且似乎还可体现“本真信息，唯一守恒”法则和“物的序位与文的序位以及可选之意的序位”的等价性或恒定性。理论上，其中蕴藏的自然法则以及逻辑和数学原理值得深究。实践上，当 $w=a+bi$ 即 z,t 为 0 时特例 $f(x,y)=w$ 有意义；当 $w=a$ 即 y,z,t 为 0 时特例 $f(x)=w$ 不仅有意义而且很管用[如：$I_U=I_D-I_K$及其相应的序位恒等式 $m_2*n_2=m*n-m_1*n_1$（体现各论域序位模型的一致性）与基于双列表分层集合的间接形式化方法结合，可支持“选域”（推理）与“定位”解（超大规模）信息方程组]。

4）“语义、信息和智”的统一理论的几何模型

语义（三棱及多棱）、信息（四面及多面）和智（四点及多点）三位一体几何模型，采用关键少数掌控多数,如：由 4 范畴到 8 体系再到 n 系列的细分（反之则是粗分乃至合一）。

5）有待探讨的问题

对广义文本多元数表达式（$a+bi\&\cdots$）和本真信息定义式（序位恒等式 $m_2*n_2=m*n-m_1*n_1$）以及智的三个发展阶段——哲学的智慧、心理学的智力、认知科学的人工智能——的探讨。

7）信息的名与实之辨

信息（概念），既不限于已有“信息”词条，也不限于是否叫“信息”。与其说“信息”是名词，不如说它是一个极为特殊的超级代词。因它几乎可指代任何“未知的或不确定的领域”。

8）本真信息的重要性及其系统表述的困难

无论知或行，支配“广义文本”的“本真信息”都是最重要的。尽管“文”形形色色、千变万化（如：文符、意念、物象），但起支配作用的却是“本真信息”（即：法、义、理、道，如：文法、意义、物理、生理、心理、哲理乃至文载之道）。对此，古今中外感悟深刻者大有人在，只因时代和知识信息数据处理能力的局限，至今未见整体上与序位模型同质的科学研究。

3. 当前信息科学交叉研究中普遍存在的问题

3.1 本体问题

徐院士采用基于哈特莱及申农而创立的公式计算本体。我采用自己的公式计算本体。我个人认为：真正能被计算的本体是“（物的）质能时空、（文的）数码、（意的）类例”的序位。

3.2 本质问题

徐院士在信源、信道、信宿的基础之上增加了信的、信值。我认为：信源、信道、信宿的研究是由“物”研究“信”，信的、信值的研究是由“意”研究“信”（这对人类具有特殊而重要的作用），数码研究是由“文”研究“信”，而“信”或信息的本质可序位化或“义”化。

3.3 现象问题

钟教授在研究信息科学原理几十年之后得到“信息好比是一个多面体”的结论。这无异于说一般信息学的研究还像是在“盲人摸象”。这一结论，应该促使人们反省以往的研究路线！我认为：可采用由“四面体”切入的方式来建立一般信息学的统一理论（框架），而具体研究“多面体”各个面的任务，主要属于部门信息学各个学科的研究任务。

3.4 名实问题

有人以是否（如：何时、何地、何人）使用过“信息”这个“词（外文）”或“辞（中文）”，作为信息学探源的依据。我认为：不妥。因为，同一个对象（物）或概念（意）或原理（义），完全可采用不同的词或辞（文）来表述。信息，也不例外。当然，研究初期这样做是必要的。

3.5 一般信息学研究的时机问题

也有人认为：在部门信息学还没有完全研究透彻之前不宜开展一般信息学研究。我认为：这对从“多面体”入手的观点来看，很有道理。但如从“四面体”入手的观点来看，一般信息学显然已有一条新的发展途径，因而，也就有了与部门信息学齐头并进协同研究的充分理由。

3.6 关于一些重要概念的语言表述问题

还有人时常谈“信息熵”，甚至有人谈“信息能”。我认为：这涉及一些知识领域的认识与语言表述的冲突问题。因为，所谓“信息熵”实质是说信息学的熵（它既不是信息也不是热力学的熵）。所谓“信息能”实质是说“信息的作用或功能”和“信息处理的能力”（其内涵区别于物理学的能和力，其外延区别于：哲学智慧、心理学智力、认知科学的人工智能）。

三、结语

我认为：要夯实一般信息学的基础，至少须有：统一的研究对象（1）、方法（2）和任务（3），即：（1）由可涵盖所有特殊信息的一般信息的存在所决定。a、哈特莱和申农实际上提出了（形式）信息（量）的统一定义。b、本文提出（形式及内容）信息（量）的统一定义。（2）由信息基本公式决定。a、bit 是（形式）信息（量）的一个常量计算单位。b、ge 将是（形式及内容）信息（量）的一个间接计算工具（相当于一系列砝码，基于分层集合的双列表就是相应的天平）。（3）由语义信息处理（智）的基本方法所决定。a、基于 bit 的数字计算技术的发展，由于具备在世界范围内达成共识乃至形成共为的理论基础和实践条件，因此取得了长足的进展。b、基于 ge 的知识信息数据处理原理、方法及实例，还在准备达成共识和形成共为的理论基础和实践条件。因此，其推广普及过程的加速或催化是必要的但要防止欲速不达。

统一的（形式及内容）信息（量）定义及其基本公式及计算单位及工具的明确，意味着：知识和语义信息的处理也可像数据处理一样得心应手。众所周知，信息科学主要是沿着形式化、数字化、数学化的可计算（形式）信息（量）bit 这一传统的直接计算途径而发展起来的，由于在语义、信息与智的基本问题上遭遇的大量歧义难以消除，所以，我在继承传统方法的基础上，另辟蹊径——提出：基于“双列表”分层集合的序位模型，其本意是为解决语言文字和知识信息的定量分析提供一套科学的量具。结果却开辟了知识和语义信息的“间接形式化”新途径。

至于提出广义信息方程，物（质、能、时、空）与信息（意、文、义）的关系，基于物的科学与基于信息的科学的关系，则是额外的发现或收获。

参考文献

1、邹晓辉“融智学初创时期的交叉研究文选(20 篇)”[C]《潜科学（前沿科学）》2005 第 48 期[EB]

2、邹晓辉“融智学精华介绍——融智学的知识创新点与基础实施例”[EB]《潜科学》2005 第 48 期[EB]

3、闫学杉“当代学科发展中的信息问题之考释”[J]《信息科学研究》2002 年第 4 期第 59-64 页

4、李宗荣“理论信息学：概念、原理与方法(博士论文)”[J]《潜科学》2005 第 49 期[EB]

5、苗东升“申论作为四论之一的信息科学”[J]《北京大学学报》（哲学社会科学版）2000 年第 6 期

6、张学文“组成论”[M]《潜科学》信息科学专栏(2005) Information Science Magazine [J] [EB]

注 1：《潜科学（前沿科学）》学术期刊网址

注 2：本文点评的几个问题是由“信息科学交叉研究学术研讨会”2005(北京) 的部分论文中发现的。

注 3：本文最后修订时采纳了闫学杉先生的意见，文章的内容和形式都得到了相应的精简。特此致谢！

特别说明：

本文不仅参考“潜科学网站”信息专栏和“前沿科学”学术期刊以及“熵、信息和复杂性网站”而且还对信息科学交叉研究学术研讨会（2005 北京）文选的论文以及“潜科学论坛”对话中本文作者发现的几个问题给予了点评。在此，向有关作者（如：邬焜、徐光宪、姜璐等）、译者（如：刘刚）和对话方（如：鲁晨光、陈雨思、金玉成等）一并表示致谢！

融智学的观点和方法

摘要：继人类智能及人工智能之后，我们发现由前两种智能结合成第三智能。其典型由自然人与计算机的合作而来。这就是为什么称融智学为研究“与计算机合作”的协同智能理论的原因。融智学由三部分组成，即：论述其合作机理的基础理论，规定其合作标准的工程技术，介绍其合作方式的教学应用。就中文信息处理在形式化的机理、标准、方式而言，协同智能无论与人类智能还是与人工智能相比都存在显著的区别。本文以三个典型示例（见：示意图 1、2、3）解释其中蕴含的新观点及新方法。

关键词：人类智能、人工智能、协同智能（第三智能）、中文信息处理

1.引言

在过去几年，人工智能的研究取得了长足的进展[1，中国人工智能进展 2003[1]；2，艾真体（Agents）[2]，知识发现和数据挖掘（Knowledge Discovery and Data Mining）[3]的研究和应用]。然而也还有很多重要的问题没有得到满意的解决[3，机器翻译[4][5]及中文信息处理与理解[6][7]的消歧困难；4，知识表达[8][9]的形式化困难；5，语义[10]、信息 [11]及智能[12][13][14][15]概念的无歧义表述的困难]。

有鉴于此，本文介绍一种协同智能的新观点及新方法，即：直接介绍 21 世纪伊始逐步公开的融智学的观点和方法，而不探讨人工智能三个流派（符号主义、连接主义、行为主义）和两种倾向（强人工智能与弱人工智能）或两种思路（由上至下与由下至上）的各种众所周知的观点和方法。

在长期探索智能现象及其原理的过程中，我们发现：人类智能主体与人工智能代理之间的合理分工、开放互动、高度协作、优势互补，实际上蕴含着一种非常重要的智能现象及原理。经多年研究、探讨和思考之后，可判定：存在一种协同智能，例如：人类智能与人工智能的组合智能。

进一步研究还发现：由自然人与计算机的合作而进化形成的组合智能只是一种非常特殊的狭义的协同智能。就像智能不限于人类一样，协同智能也不限于自然人与计算机的合作。可判定：广义的协同智能的存在。或者说：世界具有广义的协同智能的性质。真可谓：宇宙浩瀚无比，世界无奇不有，自然丰富多彩，智能无处不在。这是一种崭新的世界观和方法论，涉及博大精深的新的智能理论。其诱惑力非常大，其涉及面也很宽（由于种种限制因素，不适合此时探讨），本文仅限于介绍其中的狭义部分，旨在：抛砖引玉、窥斑知豹。希望：有提纲挈领、画龙点睛之效。

把“与计算机的合作”视为协同智能的一个典型。就中文信息处理的形式化的机理、标准、方式而言，无论是人类智能还是人工智能的优越性都不远不及协同智能。从研究该典型入手的狭义融智学由三部分组成，即：论述其合作机理的基础理论，规定其合作标准的工程技术，介绍其合作方式的教学应用。在此仅采用典型示例与精要理论结合的方式论述，可视为高级科普。

2.融智学的基础理论——理论融智学之一斑

理论融智学的精华，主要有：语义范畴新论，知识体系新论，语义信息新论，信息公理系列，信息计算原理，序位逻辑原理，表格数学[16]原理，协同智能原理，…。

2.1. 典型示例：语义三棱

在发现并指出“语义三角”不足的基础之上，我们建立了“语义三棱”模型。

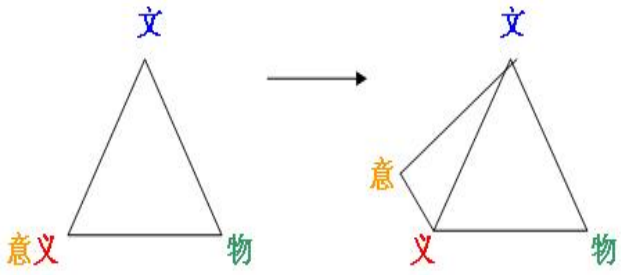


图 1 是语义三角及三棱关系示意图。

“（融智学）语义三棱”比“（语义学）语义三角”有更丰富的语义内涵。如果说“意义问题是当今人文科学（含：哲学）研究的核心问题”，那么，意与义的区分，则是（整个）科学（含：人文、自然、人工、… 科学）研究的核心问题。图 1 既可展示（融智学）一个基本的知识创新点——新的语义观，又可展示（融智学）一个根本的突破口——意与义的区分。（融智学）为什么要做这样的区分？因为我们发现：由意义到意与义的变化过程，蕴含着信息处理的原理，即：分与合的机制。同时我们还发现且构造了一个“语义三棱”，它（作为新的语义认知模型）比“语义三角”更能开拓视野。深入研究，我们由此构建了理论融智学的语义范畴新论，即：“义、文、物、意”融智概念体系。

大家知道：语义三角，是由 C.K.Ogden 和 I.A.Richards 在 1923 年出版的《意义之意义》(The Meaning of Meaning) 明确提出来的。semantic triangle（语义三角）：thing（物），meaning（意义），word or symbol（文字或符号）渊源于 1892 年 G. Frege 发表的《意义与所指》(über Sinn und Bedeutung) 明确区分“词所表达的意义和词所指的事物”的论述。哲学的语言学专向因此而开始。

由“义、文、物、意”构成的语义三棱是由融智学作者邹晓辉在 1997 年撰写并于 2000 年（在中国发明专利公报上正式）公开的“一种知识信息数据处理方法及产品（发明）”一文提出并确定的。由意义到意与义的区分和变化，凸显了中文理解语义的独特优势。这样做的必要性是这一合一分之中蕴含着信息处理机制及智能机理。新语义观的重点是：义，与之等价的概念有：法、理、道、…；难点是：意与义的区别，涉及：思想情感中的意和万事万物的理、义、法、道、…之间的区别。

融智学作者在“一个字，思考了多年——结论：信息的本质是序位本义”一文中做出了这样的解说（也许会帮助读者理解原创作者的思路）：1971-1976 期间，一个基本假设经常在我的头脑中盘旋。这个假设在 1976-1980 期间得到了进一步肯定。这就是后来一直被我称之为“信息基本定律”的那个最初的基本假设，即：同义并列，对应转换。基于该假设，可断定：“任何两个集合”，只要“同义”即可“并列”，只要“对应”即可“转换”，只要“同义并列”即可“对应转换”。后来，我发现：（1）其简单性，是它“不证自明”。于是，得出“该假设，实际上是一个公理”的结论。（2）其复杂性，在于“义”至少有“三点”（即：重点、难点、盲点）值得深入研究。重点：义，一旦得解，“信息是什么？”的问题就可能会有一般科学的唯一答案。难点：义，虽有“含义、本义、真义、…”诸多解释，但认识分歧却难消解。盲点：义，究竟可有多少解？此处它究竟指什么？似乎仍然没有直接回答。1977-1997 期间，我对“义”进行了广泛的探讨和深入的研究之后，得出一个基本判断：“同义并列，对应转换”中提及的“义”实际上是“序位本义”。

2005 年初邹晓辉在一篇国际学术会议（CLSW-6）的论文中说：这个“迟到”了 100 多年的“顿悟”告诉我们：从理论上发现“意”与“义”之间无形的“短程线”是来之不易的。可是，一旦基础理论突破之后，再由原创作者本人回过头来概括地描述它，则相对容易得多。究竟如何表述融智学的基本概念体系（即：语义范畴新论）这个深奥的原理才能做到雅俗共赏？ 经历多次尝试之后，新语义观的知识创新点及融智学通论的突破口（这一画龙点睛之笔）终于由图 1 简单的几何图形深入浅出、简明扼要、提纲挈领、恰到好处地展现了出来。图 1 呈现的意义“短程线”的发现过程是直观的。

2.2.画龙点睛：融智原理[17] [18]

理论融智学，着重论述协同智能及其与之有关的概念（如：语义，知识，信息，智能）、法则（如：信息公理，序位逻辑）和公式（如：表格数学），由通论、通则、通式“三部曲”组成，简称“三通”。

2.2.1.通论的范畴体系，由“义、文、物、意”组成，涉及：新的语义观，由“语义三棱”直观描述。如果说“通论”和“语义三棱”（模型）给出了“义”的近似解，即：序位本义，那么，“通式”中的“序位恒等式”和（工程融智学基于双列表的）“间接形式化方法”及其“(中文语汇语义信息处理和知识信息数据处理的后续两个)实施例”则给出了“义”的精确解的序位计算模型及系统。

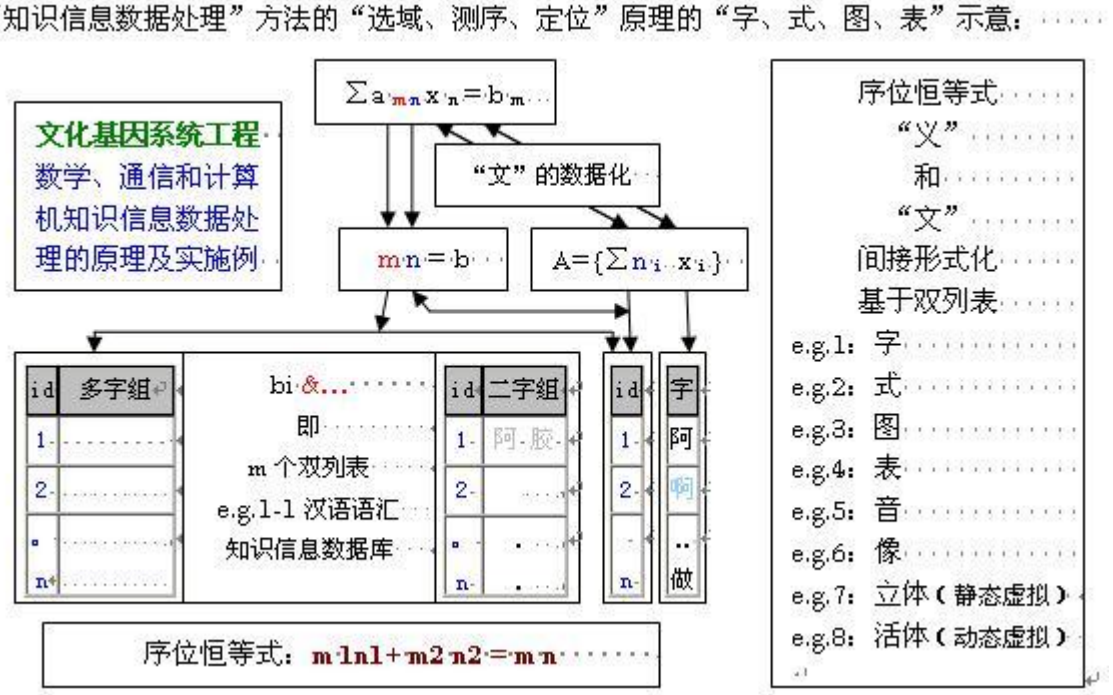
2.2.2.通则的公理系列，即：（本真信息法则 1）元素序位，唯一守恒（公理 1）。（本真信息与广义文本的关系法则 1-3）同义并列，对应转换（公理 2）；异义排列，序趣简美（公理 3）；具体层式，非

非各平（公理 4）。根据公理 1 可构建基准参照系；根据公理 2-4 可构建三个协同应对参照系。

2.2.3.通式的公式系列，即：依据公理 1-3 有 $\mathbf{m_1 n_1 + m_2 n_2 = m n}$ （本真信息法则 2 序位恒等式）。依据序位恒等式有 **m n = D** 与 **K + I = D**（数字数据公式 1 和 2）。依据数字数据公式 2 有 **I = D – K**（信息基本公式，即：狭义信息方程）。依据闵可夫斯基的时空坐标架或连续统（**x，y，z，ict**）和复数（**a + bi**）、四元数（**a+bi+cj+dk**）、… 多元数中虚数的性质，结合融智学定义的“待消歧符号：& …”（表示：不定性或多歧性）进一步定义多元数通式（**a + bi&…**）与多维可变换坐标架或连续统（**x，y，z，ict&…**）为处理最复杂的广义信息方程作必要的准备。

3. 融智学的工程实践——工程融智学之一斑

图 2 是序位数字计算原理和间接形式化方法及其实施例的综合介绍示意图。



以上图 2 的综合介绍，实际上包含：（通过中文语汇语义信息处理实施例而展示）一个基于序位恒等式及表格数字计算原理和基于双列表的间接形式化方法的全球语言定位系统（GLPS）和[通过八大形式体系的 e.g.1-8（其中 e.g.1 的一个典型实施例即中文语汇语义信息处理实施例）的前台界面、后台数据结构和算法而展示的]一个全球软件定位系统（GSPS）。

有人说：人工智能的主要的难点和重点之一是自然语言处理与理解，其中，中文语言处理与理解是难中之难，而作为其基础的中文语汇语义信息处理与理解又是重中之重。因此，工程**融智**实践的重头戏也就自然而然地放在了这个重中之重上。（体现通则和通式的）序位恒等式及表格数字计算原理和（体现通论的）间接形式化方法，是工程融智学的画龙点睛之笔。具体说明，见：图 2 及典型实例——中文语汇语义信息处理。

3.1. 典型示例：中文语汇语义信息处理

3.1.1.表格化，即：基于数字数据公式 1 和八大形式体系 *e.g.1-8* 构造的 *m* 个双列表。此处 *e.g.1* 的双列表，是：基于计算机数据库或数据仓库的电子表格（如：数据库 access 的 tables 和 SQLserver 数据仓库 cub 的 tables）。其中，左列为 *n* 个数字编号；右列为同义并列的细分进阶层式的字内文本数据及字间音节数据。由此构成两组形式化结构化计算模型：（1）处理字内信息的文本结构控制模型[STCM in word（限于 GTCM 的 0-4 进阶层式）]，（2）处理字间信息的音节结构控制模型[SSCM between words（限于 GTCM 的 4-6 进阶层式）]。其中，表格化的文本总量控制模型（GTCM）是确保自然语言得以系统地间接形式化的基础[（有关构造过程，由“中文信息处理的间接形式化新方法”具体介绍。见：参考文献）字外信息处理限于 GTCM 的 7-12 进阶层式（见参考文献，本文不讨论）]。

3.1.2.数字化, 即：m 记录表的编号，n 位于左列，记录双列表的行数及右列的格数。

3.1.3.字组化, 即：右列顺序记录。例 1：中文字内信息以笔画为基本单位而建立 STCM（字的层面型结构 1-m 细分进阶层式）。中文字间信息以字为基本单位建立 SSCM（字与字组的线串型结构 1-m 细分进阶层式）。例 2：英文字内信息以字母为基本单位建立 STCM（词的线串型结构 1-m 细分进阶层式）。英文字间信息以词为基本单位建立 SSCM（词和词组或短语的线串型结构 1-m 细分进阶层式）。

3.2.画龙点睛：基于双列表的间接形式化方法与序位恒等式及表格数字计算原理[19]

工程融智学以“三化”（即：表格化，数字化，字组化）的方式把复杂多样的对象世界简化为两个同义并列的等价模型，即：a、用**文字**描述的**间接形式化**模型[（基于双列表的 GTCM 含终极标准信息交换码 Z-ASCII）含“八大形式体系”（非纯文本部分用超级链接记于右列），即：字（如：ASCII）、式、图、表、音、像、立体、活体（如：ATGC）]；b、用**数字**描述的**直接形式化**模型[（$m_1\ n_1 + m_2\ n_2 = m\ n$）含“三类代数结构”（本文采用的代数字母所表示的数，均限于：自然数域），即：体（$a+bi\&...$）、环（$\sum a_{ij}x_j=b_i$）、群（$m\ n=b$）]，这就为应用融智学主张的融智教学参与全球知识定位系统（GKPS）共享共建和以“产、学、研、用、算”一体化为特征的融智系统工程的全面展开，奠定了形式化的坚实基础。

4. 融智学的教学应用（常规应用中的一个典型）**——应用融智学之一斑**

4.1.典型示例：信息标注

以基础和高等教育阶段的基础课课内课外教学为例，理顺：语言文字教学（以学用结合实践来检验语言文字信息标注）和通用常识教学（以学用结合实践来检验通用常识信息标注）与基于计算机辅助的生产式（语言文字和通用常识的信息标注）合作型教学活动的关系。以基础和高等教育阶段的专业课课内课外教学为例，理顺：专用知识教学（以学用结合实践来检验专用知识信息标注）与基于计算机辅助的生产式（专用知识信息标注）合作型教学的关系。由现行教材、教参和课件作为起步的基础。

4.2.画龙点睛：融智教学

面对（学习、生产、生活中经常遭遇的）“三多”（即：多物象、多文符、多歧义）现实情景，（工程融智学）“三化”（即：表格化，数字化，字组化）和（应用融智学）“三注”[即：（基于广大师生经常性学用结合实践检验的）语言文字、通用常识、专用知识的三级标注]的知识信息数据加工制度及程序（以融智教学的方式在“初试、中试、商品化”的良性循环过程中不断地发展完善）是全域数码定位系统[如：基于“三化”的 *GLPS*（含：*GTCM；STCM，SSCM* 或 *GSCM*），*GSPS*（含：*Z-ASCII*），基于“三注”的 *GKPS*]的两个基础，即：工程**融智**和应用**融智**。通常情况，工程在前，应用在后。实际上，这只是静态的观点。如果从动态的观点看，处于动态循环过程中的前后关系是不断变化的。正是充分考虑到了动态发展的循环因素，才特意设计了一套有利于动态循环发展的融智教学模式（即：基于协同智能的生产式合作型教学模式是应用融智学的画龙点睛之笔），用以支持工程融智学设计初创的“全域数码定位系统”的优化发展。

图 3 是计算机辅助“三注”的知识信息数据库的生产式合作型教学共享共建全域平台示意图。

以上图 3 呈现的一个基于 *GLPS* 的 *GKPS* 共享共建全域平台，是：一个以“产、学、研、用、算”有机结合为特征的生产式合作型教学融智过程（即：融智系统工程的一个重要方面）的标准化与个性化结合的开放式进化发展平台，其作用：一方面是方便计算机辅助教学；另一方面是方便广大师生与遍布各级学校乃至校外的网络化计算机系统之间“产、学、研、用、算”一体化的大协作、大融通、大融合（将彻底改变各个教师乃至各专家小组在教学、教研、科研上闭门造车的常规化方式）。

5.结语

综上所述，由理论融智学、工程融智学、应用融智学三部分组成的融智学是崭新的学问体系（即：理论与实践互动的融智三部曲）。其中，理论融智学研究协同智能的存在机理，工程融智学研究狭义协同智能的形成机制及构造方法，应用融智学研究狭义协同智能活动的组织方式及管理模式。如果说**工程**融智学是**以“三化”的方式**把**复杂多样的对象世界**简化为**两个同义并列的等价模型**，那么，**应用**融智学则是**以“三注”的方式**[把千变万化的对象（典、例、类）的超级链接页面及标注查询列表]**添加或填充到**基于上述第一个等价模型而建立的知识信息数据库的标注查询列表中并以教研及教学实践不断地检验它，从而，形成共享共建的标准化智能平台与个性化智能代理有机结合、和平共处的知识信息数据处理协同智能系统。随着融智教学活动的不断推进，用户（无论设计者还是使用者其知识信息都不断得到扩充）与系统（其应用范围也不断得到扩展）。

通过人类智能与人工智能的合理分工、开放互动、高度协作、优势互补而形成狭义的协同智能，一方面，可开辟一条坦途以解决机器翻译及中文信息处理与理解等人工智能领域举世公认的消歧难题。另一方面，可建构一种在人与人、人与机、机与机、机与人之间进行大协作、大融通、大融合的“产、学、研、用、算”一体化的融智活动方式。其独特之处在于：首先，在工程融智阶段，采用：基于双列表的间接形式化方略，其基础是在理论融智学的基本观点和方法的指导下，综合：语言学、认知科学、逻辑学、数学、通信和计算机技术以及计算语言学（同时特别吸取了其中机器翻译及中文信息处理与理解的经验教训）等多学科知识技能而形成的一套融智系统工程方法；接着，在应用融智阶段，把人类智能与人工智能相结合而形成的狭义的协同智能活动融入与工程融智实践的互动过程之中，为构建“产、学、研、用、算”一体化融智系统工程提供一个标准化与个性化结合的广阔发展新境界和美好前景。最后，在理论融智阶段，还可再进一步，即：不仅可定义继人类智能及人工智能之后出现在自然人和计算机之间的狭义的协同智能，而且，（在全域数码定位系统和生产式合作型教学实践的支持的基础之上）还可定义宏观及微观世界具有更加广阔探讨价值的广义的协同智能。随着融智理念的普及，必将显著地拓宽人们对智能的认识视野和实践领域。

除了以上重点介绍的融智学方法概要之外，下面再举要介绍几个融智学的新观点：

（1）新的语义观：主张用四范畴发展三范畴（即：物，如：事物；文，如：符号；意义，如：概念）。

四范畴，即：义，指：本真信息；文，指：符号形象；物，指：载体载能；意，指：意识意向。举例说明：原理，如：杯子的机理，是：本真信息，属于：义的范畴；展示其机理的文化形式，如：杯子的图纸，属于：文的范畴；展示其机理的物化形式，如：具体的杯子，属于：物的范畴；智能主体的选择，如：杯子的构造及外观的设计构想，属于：意的范畴。

（2）新的信息观：首先，主张用精准描述的三要素（质、能、信）取代含混描述的三要素（物质、能量、信息）；然后，主张用三范畴（义，文，意）细化一要素 [（通信与信息的）信]。

（3）新的知识观：主张用“语义三棱模型”定性地宏观把握具体而粗放的“意”（即：对“义，文，物，意”做具有明确范畴方向的选择）和以“三化”及“三注”构成的“全域数码定位系统”定性且定量地微观掌握具体而精准的“意”[，即：一系列选择的记录（即：对“三化”及“三注”做具体精确而可重用的选择）]来发展或取代一要素（即：意义）粗放而含混的“知识”（如：概念）。

本文是一篇随机选点介绍融智学探索成果的短文，不能像（正在撰写中的）专著那样具体周到。

参考文献

[1] 中国人工智能学会：中国人工智能进展[C]北京邮电大学出版社 2003

[2] Software Agents[C]ISBN 0-262-52234-9 AAAI Press

[3] Advances in Knowledge Discovery and Data Mining[C]ISBN 0-262-56097-6 AAAI Press

[4] 陈肇雄主编：机器翻译研究进展[C] 1-564 页，电子工业出版社，1992

[5] 黄河燕主编：《机器翻译研究进展[C] 1-282 页，电子工业出版社，2002

[6] 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

[7] Recent Advancement In Chinese Lexical Semantics [A] CLSW-5 [C] Singapore,2004

[8] 邹晓辉：一种知识信息数据处理方法及产品[J]发明专利公报 G06F163 知识产权出版社，2000，（11）

[9] 邹晓辉：协同智能计算语言数据库的设计方法[J]潜科学（第 32 期）2004（7）

[10] 邹晓辉：协同智能计算知识数据库的设计方法[J]潜科学（第 39 期）2005（1）

[11] 邹晓辉：优化“语义信息处理”的新方法与实施例（间接形式化方法）[A] CLSW-6 [C] 厦门大学 2005

[12] 邹晓辉：重构“概念分类体系”的新思路与新方法（语义三棱模型）[A] CLSW-6 [C] 厦门大学 2005

[13] 邹晓辉：中文信息处理的新方法（间接形式化) JSCL-2005 [J]潜科学（第 42 期）2005（4）

[14] 邹晓辉：默契通信与间接计算对自然语言处理的重要性 JSCL-2005 [J]潜科学（第 42 期）2005（4）

[15] 邹晓辉：语义信息新论（信息基本公式）ISM 信息科学研讨会-2005[J]潜科学（第 43 期）2005（5）

[16] 张学文：组成论[M] 44-56 页，（字符多项式与表格数学）246-252 页，中国科学技术大学出版社 2003

[17] 邹晓辉：广义文本与序位本义（本真信息）ISM-2005[J]潜科学（第 43 期）2005（5）

[18] 邹顺鹏、邹晓辉：两个基本信息公式及其算法的坏与好的比较[J]潜科学（第 44 期）2005（6）

[19] 熊全淹：近世代数[M] 15-120 页，上海科学技术出版社，1978

尾注

本文写作的过程中还参考了以下网络文献：

AI in the news ©2000 - 2005

中国人工智能学会第 10 届全国学术年会论文集

陆汝钤：发展知识工程 建立知识产业

潜科学 2002-2005 各期有关信息、知识、智能的探讨和融智学的文章

Cooperative Learning Methods

后　　语

融智学作为一个崭新的领域，其主旨是探讨继人类智能和人工智能之后而出现的协同智能，其典型是自然人与计算机的合作（因此狭义的融智学可译为“与计算机合作”的智能理论），由三部分组成，即：论述其合作机理的基础理论，规定其合作标准的工程技术，介绍其合作方式的教学应用和经济应用乃至其他领域的应用。

作者可在本书中直接介绍融智学，要感谢中国人工智能学会 2005 年全国学术大会（CAAI-11）的专家们！他们对“融智学的观点和方法”（高级科普文章）、“自然语言处理的总量控制模型---形式化标准平台”和“理性的标准的协同智能模型”三篇科学论文的认可，坚定了作者把融智学直接用于探讨“字本位与中文信息处理的基础”这一课题或难题的决心。

与“字本位与中文信息处理”有关的文章占据了本书的大部分，在选题初期，作者得到了“全国普通高等学校人文社会科学研究十五规划纲要”语言学咨询组负责人徐通锵教授的直接指点和帮助。在此文集出版之际，请允许笔者对徐老表示由衷的敬意和谢意！没有徐老引导就没有本书的这个部分，作者也不会在基础语言学领域思考这么深入。

作者还要感谢鲁川教授（中文信息学会首届计算语言学专业委员会主任）和苑春法教授（清华大学国家智能实验室自然语言处理与理解专家）！没有鲁老这一“知音”[鲁老认为：工程融智学间接形式化体系的“13 张表的构建充分体现出你（指：笔者）能站在一个较高的起点上善于集中现有各家学派的优点”，苑教授认为“这 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。”2002，11]，没有他们的支持和鼓励，笔者也不会这么有信心。

同时，作者还要感谢易绵竹教授（解放军外国语学院计算语言学研究室主任，国际信息化科学院院士）！应当指出：融智学推广初期，作者与易教授的学术交流具有特别的作用或意义。当时如果没有易教授的鼓励和提示，那么，作者也许还不知道自己的融智学探讨竟然会对计算语言学（自然语言处理与理解）产生如此巨大的影响[易教授认为“当前语义研究的理论方法还需融合统一，您（指：本书作者）创立的融智学新范式提炼出协同智能主体的概念体系具有原创性，想必对自然语言语义信息的处理将引发一场革命。”2001-09-25]。

作者也要感谢期盼着本书早日问世的学界同行和众多的友人[他（她）们至少涉及：数学、语言学、计算机科学、软件工程、知识工程、人工智能、机器翻译、计算语言学、中文信息处理、系统科学、信息科学、融智学、英语教学、大学教务及研究生管理等不同科学领域）以及[作者读研究生期间的母校吉林大学和明确表示支持作者推广融智学成果的清华科技园（珠海）]领导及同仁，他（她）们（按照汉语拼音字母表由计算机自动排序）是：曹存根、常宝宝、陈群秀、陈雨思、陈肇雄、董振东、冯志伟、关培忠、郭雷(中科院院士)、胡俊锋、黄昌宁、黄河燕、黄建华、黄曾阳、姬东鸿(新加坡)、李素元、林杏光*、鲁晨光、鲁川、陆俭明、陆汝黔(中科院院士)、马喜腾、孟华、邱嘉文、求式纶、求同格、宋福群、苏新春、孙怀庆、孙茂松、王洪君、王惠(新加坡)、王源良、奚建清、徐德华、徐通锵、闫学杉、杨自俭、易绵竹、于江生、俞士汶、苑春法、詹卫东、张拔(中科院院士)、张普、张全、张学文、钟义信、周明、周强、周新全、…

没有与他（她）们的交流或他（她）们大大小小的鼓励或激励及各种各样的支持或帮助，作者不会如此深入地研究这一涉及面如此广泛的难题而仍能保持如此巨大的热情和勇气。

最后，致谢家人

邹晓辉 2006-3-25 于恒美花园

2025 年第三次再版摘要

《融智学原创文集》汇集了作者 2000-2005 年间在学术期刊和会议上发表的文章。尽管作者的认识已深刻且系统化，但考虑到文集的学术交流价值，特别是其中保留的对多学科交叉问题的原创成果和探讨，对跨学科研究人员有借鉴、启迪或警示作用。文集保留了原创成果初创时期的基本风貌，具有创新知识点和广泛的社会经济实用价值。欢迎读者提出宝贵意见，希望就新兴学科与周边交叉学科关系的问题进行科学探讨，特别希望能与关注“协同智能计算系统”和“融智与融资”的读者交换意见。**该书 2007 年第一版，2018 年第二次再版（ISBN:**

9780463607640），2025 年第三次再版（邹晓辉 2025-1-24 于横琴粤澳深度合作区仁山路 100 号 1 栋）。

"The Original Essays on Integrated Intelligence or Original Collection on Smart-System Studied" is a collection of articles published by the author in academic journals and conferences from 2000 to 2005. Although the author's understanding has become more profound and systematic, considering the academic exchange value of the collection, especially the original achievements and discussions on multidisciplinary intersection issues it preserves, it serves as a reference, inspiration, or warning for interdisciplinary researchers. The collection retains the basic features of the original achievements during their initial stage and possesses innovative knowledge points with extensive socio-economic practical value. Readers are welcome to provide valuable feedback. We hope to engage in scientific discussions with readers on how emerging disciplines handle relationships with surrounding interdisciplinary fields, and particularly look forward to exchanging views with readers interested in "Collaborative Intelligent Computing Systems" and the dual practice of "Integrated Intelligence and Financing". The book was **reprinted for the first time in 2007** and for the second time in 2018 （ISBN: 9780463607640）and for the third time in 2025.

本文集作为融智学导论的《字本位与中文信息处理的基础》和作为融智学三部曲的《理论融智学》、《工程融智学》与《应用融智学》几部专著的形成奠定了相应的基础。