国际汉语教学学报（2024）
第 05 卷 · 第 02 期 86-93 https://doi.org/10.46451/ijclt.20240208

百宝箱语素字词典

夏诠真

摘要
直到目前,汉语只有字被编码。无论是手机书写或是网上阅读,所有的文章,全部是以 unicode（国际统一码）为内存。Unicode 是字形的代码，然而，文章其实是句的组合，句是词的组合，词是最小的理解单位。以 unicode 去记忆汉语文章是间接方式，unicode 不带义，所以无论是人阅读或是机器翻译都会被误解或曲解。百宝箱的思路是改用概念码（词的代号）来记忆文章，概念码是用语素定义，于是文字变得非常清晰准确，可以帮助汉语进入人工智能领域。本文详细解释语素码的设计。

关键词
统一码，概念码，中性码，语素码

1 理论基础

汉语须被编码然后才能进入计算机应用。无论是英语或是汉语,文章是句组成,句是词（广义）组成。英语词是“字母”组成；汉语词是“中文字”组成。中文字多音多义，以“字”组词的负面是计算机无法辨识其读音和含义。

百宝箱的思路是改用概念码为中文的内存，文章不再是字的组合而是概念码的组合。由于概念码已经很准确被中性码和语素码所定义，所以改变内码之后，汉语文章的清晰性和准确性被大大地地改善。

本文解释概念码和语素码的核心设计。

2 百宝箱的语素码与语素处理

2.1 百宝箱的语素码概念

汉语文章是由句组成，句是由词（广义，包括成语、俗语、惯用语等短语）组成，词是字的组合。汉字多义，以字去定义词不够清晰，会被误解或曲解，所以百宝箱改由语素码为词定义。什么是语素码? 汉语的语素来自英语的 morpheme 观念，但是由于汉语的特殊性，它被“汉语化”，比英语的 morpheme 更丰富和多用途。百度百科 ' 说 :『语素是一个语言学的新术语，指语言中最小的音义结合体。一个语言单位必须同时满足三个条件——“最小、有音、有义”才能被称作语素。』

联系电邮：rene_he@aol.fr

百宝箱尊重语言学的观点去定义和制作语素码，除了满足最小、有音、有义三个条件之外，我们再补充一个规定：“语素”只有一个含义，但是为了简化处理逻辑（以语素为词定义），百宝箱的语素码可以是几个邻近语素的合并（譬如一个字的基本义加其引申义，或数个含义相近的语素）。必须遵守的原则是：①若字有数个明显的不同意思，每个意思制作一个语素码；②若字有数个读音，每个读音制作一个语素码；③若字有数个繁体字形，每个字形制作一个语素码。

(1) 字多音的例子：“行”可以读 xíng 或读 háng。

(2) 字多繁体形的例子：简体“历”对应繁体“歷”和“曆”字。

(3) 语素合并例子：“日”字是一个多义字，它有下述多个含义：

- 离地球最近的恒星（亦称“太阳”）：日月星辰。
- 白天，与“夜”相对：日班。
- 天，一昼夜：多日不见、今日、日程。
- 某一天：纪念日。
- 计算工作的时间单位，通常以八小时或六小时为一个工作日。
- 时候：春日、往日。
- 每天，一天一天地：日记、日益。
- 特指“日本国的”：日元、日货。

百宝箱将第二至第七个含义合并，制作 0112A（太阳）、0112B（天数或白天）和 0112C（日本国的）三个语素码。

语素有单音节和多音节的划分。百度百科说：

『双音节语素，组成该语素的两个音节合起来才有意思。双音节语素主要包括联绵字（能再细分为双声、叠韵、非双声叠韵联绵字三类）、外来词和专用名词。多音节语素由两个音节以上的语素组成；主要是拟声词、专用名词和音译外来词。例如：喜马拉雅、珠穆朗玛、安迪斯、法兰克福、奥林匹克、白兰地、凡士林、噼里啪啦、淅淅沥沥、马克思主义、中华人民共和国、迪斯科、尼古丁、乌鲁木齐、布尔什维克、唏里哗啦……。』

备注 1: A. 双声是声母相同的联绵字: 如琵琶、乒乓、澎湃、鞑靼、尴尬、荆棘、蜘蛛、踯躅、踌躇、仿佛、瓜葛、忐忑、淘汰、饕餮、倜傥、含糊、慷慨、叮当、蹊跷、玲珑、犹豫等。B. 叠韵是韵母相同的联绵字: 如从容、葱茏、葫芦、糊涂、匍匐、灿烂、蜿蜒、苍茫、朦胧、苍莽、邋遢、啰嗦、怂恿、螳螂、枞椤、倥侗、蜻蜓、轰隆、当啷、恼恍、魍魉、缥缈、飘渺、奋拉、实施等。C. 非双声叠韵联绵字: 如蜈蚣、蓊郁、珊瑚、疙瘩、蚯蚓、惺忪、铃铛、奚落、褡裢、茉莉、蚂蜡、窟窿、伉俪、蝴蝶、笊篱、蹦达、蟪蛄、狡狯、狡猾、蛤蚧、蛤蜊、牡丹、磅礴、提溜等。

备注 2: 外来词是由汉语以外的其他语种音译过来的词语。如干部、涤纶、夹克、的士、巴士、尼龙、吉普、坦克、芭蕾、哒爹等。

备注 3：专用名词主要是地名和人名。如纽约、巴黎、北京、苏轼、李白、孔子。

2.2 百宝箱的语素处理

由于语素的主要功能是作为构成词和短语的材料，所以我们是从词的分析角度去制作语素码。百宝箱的《现代汉语常用词表》拥有五万多个词，我们于是统计这五万多个词的“字”的出现次数，按照字的出现频率制作了一个《规范字常用度表》，显示“字”的基本属性（繁体形、拼音、部首、声符、笔画数、组词例子），将出现次数多的字（子、不、人、大、一、生、心、水、头、无、儿、风、气、天……）排在最先。

备注：汉语规范字的数量是 8105 个，但其中只有 5100 个被应用于组词。

在《规范字常用度表》的基础上，我们再分析《首选语素码表》制作《中性码 - 缺省语素码表》和《常用字语素分析表》。第一个表显示字和中性码之间的从属关系，第二和第三个表显示字、中性码和语素码之间的协作关系，每个语素码都注明其含义和别的属性。

下面是这三个表的内容和结构：

表 1. 规范字常用度表

表 2. 中性码 - 缺省语素码表

表 3. 常用字语素分析表


常用字语素分析表

汉典 查查 百度

规范字	中性码	繁体字	拼音	语素	语素义	构词例子
子	0064V	子	zǐ	0064A	①古代指儿女，现专指儿子。②与“母”相对。	①子女，子孙，子嗣，子弟（后辈人，年轻人）。②子金（利息），子母扣，子音（辅音）。
	0064V	子	zǐ	0064B	①植物的果实、种子。②动物的卵。③幼小的。④小而硬的颗粒状的东西。	①菜子，瓜子儿，子实。②鱼子，蚕子。③子鸡，子畜，子城。④子弹，棋子儿。
	0064V	子	zǐ	0064C	①对人的称呼。②古代对人的尊称；称老师或称有道德、有学问的人。	①男子，妻子，士子（读书人），舟子（船夫），才子。②孔子，先秦诸子。
	0064V	子	zǐ	0064D	①地支的第一位，属鼠。②用于计时。	①子丑寅卯（喻有条不紊的层次或事物的条理）。②子时（夜十一点至一点），子夜（深夜）。
	0064V	子	zǐ	0064E	封建制度五等爵位的第四等。	子爵。
	0064V	子	zǐ	0064F	①附加在名词、动词、形容词后，具有名词性（读轻声）。②个别量词后缀（读轻声）。	①旗子，乱子，胖子。②敲了两下子门。
	0064W	子	zi	0064G	词尾。	

计数 996

审校人 杨兰

审校日 05/12/2023

Y/N Y

建议

2.3 如何以语素码为词定义？

《规范字常用度表》、《中性码 - 缺省语素码表》和《常用字语素分析表》这三个表的制作都是为了以语素码为词定义，将“词 = 字 + 字”结构改变为“概念码 = 语素码 + 语素码”结构，改善汉语的表达清晰性和准确性。

字和语素是数学“一对多”的关系，一个字对应多个语素。从“字”结构转变为“语素码”结构是艰巨的工程。《百宝箱常用词表》有五万多个词，为五万多个词以语素定义工作量很大，必须使用“人机合作”的方式去做。我建议的处理工序是：①中性码表的制作；②语素码表的制作；③常用词的处理；④以中性码为概念码定义；⑤以语素码为概念码定义。

2.4 中性码表的制作

汉字多音多形（正、简、繁、异），为了表音和简繁文本的准确互译，百宝箱制作了《中性码表》，它的基本字段是：中性码、拼音、规范字形、标准繁体字形、常见异体字形、首选语素（首选语素的应用见后）；此外，每个规范简体字形和标准繁体字形都各自记忆了它们的部首、声符、笔画数、笔顺和 unicode 号码。

例子：“干”字是一个多形多音字，它对应干、乾、幹三个繁体字形：

- 繁体“干”- 音 g ā n，部首“干”，声符“干”，笔画数“3”，笔顺“112”，首选语素“0023A”，Unicode = 5E72；
- 繁体“乾”- 音 g ā n，部首“乙”，声符“幹左”，笔画数“11”，笔顺“12251112315”，首选语素“0023D”，Unicode = 4E7E；
- 繁体“幹”- 音 gàn，部首“乙”，声符“幹左”，笔画数“13”，笔顺“1225111234112”，首选语素“0023G”，Unicode = 5E79。

2.5 语素码表的制作

汉字多义，为了清晰表意，百宝箱制作了《语素码表》，它的基本字段是语素码、中性码、拼音、短释义、定义或注解、例句、英语。汉字有些含义相近，有些含义极少用。我们采取务

实方针去制作《语素码表》：①无义表音字（譬如拟声字、音译外来词）不制作语素码；②意思相近的字义（譬如“月”字有 a. 计时单位和 b. 按月出现的两义）合并成一个语素码；③罕见的字义（譬如“厂”的一个罕义是“山边岩石突出覆盖处”）不制作语素码。制作或不制作语素码的原则取决于语素的组词能力。有很多字不出现在常用词中，这些没有组词能力的字我们不制作语素码。

例子：“面”是一个多繁体形和多义字，它对应下述四个语素码：

- 1455A – 中性码“1455V”，繁体形“面”，音 miàn，定义是：①头的前部，脸；②用脸对着，向着；③会见，直接接头的；④和颜面有关的，构词例子：①脸面、颜面、面目、面面相觑；②面对、面壁；③当面、面议、面晤、耳提面命。
- 1455B – 中性码“1455V”，繁体形“面”，音 miàn，定义是：①事物的外表；②几何学上指线移动所生成的形迹，有长有宽没有厚的形；③量词，多用于扁平的物件，构词例子：①地面、面友、面额；②平面、曲面；③一面鼓。
- 1455C – 中性码“1455V”，繁体形“面”，音 miàn，定义是：方位，部分；构词例子：前面、反面、片面、全面、多面手。
- 1455D – 中性码“1455W”，繁体形“麵”，音 miàn，定义是：①粮食磨成的粉，特指小麦磨成的粉；②由面粉和水做成的条状食物；③粉末；④食物含纤维少而柔软；构词例子：①面粉、面食、面包；②面条；③药面儿；④这种瓜很面。

3 常用词、多音字与多义字的处理

3.1 常用词的处理

中文有不少单音节词（天、地、中、长、行、走、里、面……），单音节词一般多含义；有些多音节词（出口、方便、交通、意思、粉丝……）也多义。词语的词性、读音、英语对应词是跟从词义的，百宝箱于是将多义词拆分成多个概念码，概念码的基本字段是：概念码、简体字形、繁体字形、字数、拼音、定义、例句、英语对应词、法语对应词、近义词、反义词、词类号码、词结构。词的结构是：单纯词（由一个单音节或一个多音节语素定义的词）、复合词（词根＋词根）、派生词（前缀＋词根；词根＋后缀）。

将《百宝箱常用词表》转换为《百宝箱常用词概念码表》后，我们制作两个常用词副表：A. 须以语素码定义的词（主要收集复合词和派生词）；B. 不须以语素码定义的词（主要是联绵词、拟声词、外来译音词、构词后原字意消失的词、专有名词、专科术语）。

为什么有些词不需要以语素码定义？无论是联绵词（踯躅、踌躇、螳螂……）、拟声词（噼里啪啦、滴答滴答、叮叮咚咚……）、外来译音词（引擎、和尚、盘尼西林、安培……）或专有名词（李白韩愈等人名、福建深圳等地名……），组成词的单字只用来表音，没有含义，所以不需要为它们以语素定义。有些词以字组合之后，原字义消失或不显著，所以也不需要为它们以语素定义。

原字义消失或不显著的例子：“开心”的意思是心情愉快舒畅，组成“开心”的“开”字，其字义和“开心”的语义毫无关系。同样，“方便”的“方”字，“打的”的“打”字，其语素不显著。百宝箱系统于是不为开心、方便、打的、物理、会计、金融、交通、共和……这些词以语素码定义。它们被收集进 B 副表，当做“多音节语素码”去处理。

3.2 以中性码为概念码定义

字多音（长、行、差、藏、区、校……），有些字（规范字）可以对应多个繁体字（面、里、历、发、须、了、几……），因此以字为词定义使汉语的电子化处理变得复杂易错。百宝箱系统的解决方式是改以中性码为词定义。字和中性码的关系是一对多，从“词＝字＋字”结构转变为“词＝中性码＋中性码”结构并不容易。我们的处理程序是：①使用 access 的内建 SQL 程序将全部五万多个常用词拆分成“词＝繁体字＋繁体字”结构；②凭《中性码表》，找繁体字的中性码（以 SQL 程序做）；由于一个繁体字可以对应数个中性码（是一对多的关系），所以我们再以人机合作的方式，凭词的拼音改正自动处理的错误；③人力审改《中性码表》；④使用程序为每个中性码制作首选语素码（语素的缺省值）。

举例，中性码 0006V“厂”对应 0006A（工场、棚舍）和 0006B（中国明代设的特务机关）两个语素码，我们以 0006A 为首选语素码（因为 0006A 的组词频率比 0006B 大）。中性码 0146V“分”对应 0046A（区划开、散、离、辨别）、0046B（一半）和 0046C（计算成绩的单位），我们以 0046A 为首选语素码。

3.3 以语素码为概念码定义

字多义（里、打、下、分、点、卜、黑、牛、方、把、校……），以字为词定义妨碍汉语的进一步电子化。百宝箱系统的基本设计是使用概念码为文章的内码，每个概念码再以中性码和语素码定义。以语素码为词定义是非常困难的事情，我们的解决方法是从现成的“概念码＝中性码＋中性码”结构转变为“概念码＝语素码＋语素码”结构，处理工序是：①审查常用词 A 副表（须以语素码定义的词），使用 SQL 程序阅读《中性码表》找寻“首选语素码”，将“词＝字＋字”结构转变为“概念码＝首选语素码＋首选语素码”结构；②组织多个审校工作小组，人工修改自动处理的错误（按照词义，将一些首选语素码改掉）。

语素码定义例子：“中”是一个多义字，它有六个不同含义，百宝箱系统为它制作了六个语素码：

（1）0113A（音 zh ō ng）- 意思是：①和四方、上下或两端距离同等的地位；②在一定范围内，里面；③性质或等级在两端之间的。

（2）0113B（音 zh ō ng）- 意思是：表示动作正在进行。

（3）0113C（音 zh ō ng）- 意思是：特指“中国”。

（4）0113D（音 zh ō ng）- 意思是：适于，合于。

（5）0113E（音 zhòng）- 意思是：①恰好合上；②科举考试被录取。

（6）0113F（音 zhòng）- 意思是：受到，遭受。

中性码“中”（0113V）的首选语素码是“0113A”，于是自动处理将全部“中”语素码都定义为“0113A”（英语 middle）。然而，中华、中餐、中文、中医……这些词中，“中”的意思是“中国的”，英语 Chinese，因此必须改用“0113C”为语素码。

4 制作《百宝箱语素字词典》的用意

我们起意制作《百宝箱语素字词典》的动机是：

（1）中国的工具书品种多，汗牛充栋，但是还没有语素字词典;制作《百宝箱语素字词典》填补这个缺漏。

（2）《汉语百宝箱方案》是十年长，投资额庞大的计划。长安城非一日可建，我们须将百宝箱方案拆分成多个中、小项目去一步一步完成。《百宝箱语素字词典》是汉语百宝箱方案中的一个中型项目，投资额是 3000-5000 工时左右，估计一年至二年后可以完成，是百宝箱协会力之所及。

备注：与大学及 / 或企业合作是我们的愿望，但是百宝箱团队不须等待，可以立即动工，担任先头部队。

（3）用概念码（词的代号）替代 unicode（字的代号）为文章的内存是汉语电子化路程上的一项重大突破性改革，制作概念码任重道远，工作量非常大。俗语说：积习难返，虽然概念码的优点很多，但是由于人的惰性，它的推动必然会遇到抗拒，很多人质疑以词的代号替代字的代号的可行性和需要性。我们是想以《百宝箱语素字词典》的实用性和优越性来证明汉语百宝箱计划并非空中楼阁。

5 百宝箱语素字词典的特色

《百宝箱语素字词典》和新华、现代汉语等印刷版字词典及汉典、百度百科等网上查询工具非常不同，它的特色是：

（1）以关系数据库储存信息。关系数据库的优点是它按数据的类别性质，以多个独立但互相链接的表格记忆信息，非常适合用于字词典的构建；

（2）以《中性码表》记忆和显示“字”的读音和正、简、繁、异体字形的属性（部首、声符、笔画、笔顺、unicode）。用户可以凭部首、声符、笔画、笔顺（及它们的组合）去找字，迅速方便。

（3）以《单音节语素表》显示“字”的含义和组词例子（“字”的每个基本含义以一个语素码记忆；其中一些语素码对应英语词）。

（4）以《多音节语素表》显示单纯词的含义和组词例子（“天文”是一个单纯词，它可以组成：天文表、天文单位、天文馆、天文数字、天文台、天文望远镜、天文学、天文钟等合成词）。

（5）以《规范字字形表》显示所有规范字（正、简、繁、异）的字形属性（部首、声符、笔画、笔顺、unicode）。这个表的主要用途是凭字形属性迅速搜索到需要的字。

（6）文本（句的组合）不是用字来定义而是用“概念码”来定义;概念码有唯一性（多义词，每个义制作一个码）。

（7）概念码都中英对照（也有法语），面向世界（外语能帮助理解词的准确含义）。

（8）概念码是使用中性码和语素码来定义，方便简繁字形的转换和允许用户做精准的信息搜索、统计和分析。

（9）所有的百宝箱表格都拥有排序、检索、筛选、分类、统计……多种内建功能。

（10）新华、汉典……等字典主要显示“字”属性和解释字的含义，以“字”为中心；百宝箱语素字典显示“字”和“词”之间的唇齿主从关系，主要用于帮助用户修辞，做细腻的分析，以“语义”为中心。

（11）一般字词典主要是用来查字的解释，于是收集了大量的艰涩字词；《百宝箱语素字词典》是从实用角度（帮助用户改善修辞）去设计，只收集最通用的字词。

6. 百宝箱语素字词典的构建部署

《百宝箱语素字词典》的建设程序是：

（1）调查教师们（语素字词典的最大用户群）的需要和愿望，拟定目标，规划数据库的内容和功能；

（2）估计工作量和人力需求；拟定预算（资金数目，工作进度、完工期）；

（3）成立管理机构；组织多个工作小组（汉语小组、英语法语小组、审校小组），分工合作；

（4）制作试用版，由教师们测试，回馈心得；

（5）完成《百宝箱语素字词典》的个人电脑版（使用微软的 ACCESS 关系数据库）；

（6）考虑是否开发《百宝箱语素字词典》的网络版及 / 或印刷版。

7. 百宝箱语素字词典的商业营运

百宝箱协会是公益组织，没有谋利企图，我们只投入能调动的人力和财力去尽心制作《百宝箱语素字词典》数据库，将它的商业营运（制作能赚钱的实用产品）付托给愿意投资和肯承担风险（自负盈亏）的企业（譬如一家出版社），由它全权去推动和经营。

8 结语

Unicode 已经被应用了三十年，它很适合表达全球文字，但是在非常复杂的汉语方面，它不适应，妨碍汉语的机械化处理。百宝箱为词编码是结构性改革，一旦完成，中文的地位能与拉丁语系平起平坐。

参考文献

1. https://baike.baidu.com/item/%E8%AF%AD%E7%B4%A0?fromModule=lemma_search-box

投稿：2022 年 8 月 17 日；接受：2023 年 9 月 18 日；出版：2024 年 2 月 21 日

作者简介

夏诠真在越南和柬埔寨长大，1967 年到法国格勒市攻读信息学及工商管理，先后于 Westinghouse，Sisco，UTA，Air-France 等大公司担任工程师和项目经理之职，开发了无数大型软件系统。1998 年退休后投身于中文的编码研究，聚合一班优秀年轻人共同推动汉语百宝箱计划。

International Journal of Chinese Language Teaching (2024)
Vol. 5 (2) 86-93 https://doi.org/10.46451/ijclt.20240208

The Morpheme Dictionary of the Chinese Toolbox

Qianzhen Xia

Abstract
Up to now, only individual Chinese characters have been encoded. Whether it's writing on mobile phones or reading online, all articles are stored using Unicode (globally unified code). However, articles are composition of sentences, and sentences combinations of words, while words are the smallest units of meaning. Remembering Chinese texts by using Unicode is an indirect method, as Unicode does not bear meaning, thereby resulting in misunderstanding or twisted meaning, no matter whether it is translated by machine or read by human beings. The concept proposed by the Chinese Language Toolbox project is to use conceptual codes (meaning of words) to replace Unicode. The conceptual code is defined by morphemes, making Chinese characters clearer and more accurate and helping the Chinese language into the artificial intelligence era. This text provides a detailed explanation of our design.

Keywords
Unicode, conceptual code, neutral code, morpheme code

Mr Quanzhen Xia (Rene HE) grew up in Vietnam and Cambodia. In 1967, he went to Grenoble, France, to study Information Technology and Business Administration. He subsequently worked as an engineer and project manager at major companies such as Westinghouse, Sisco, UTA, and Air France, contributing to the development of numerous large-scale software systems. After retiring in 1998, he has been devoted to the research of Chinese coding and has gathered a group of talented students to work together for the Chinese Language Toolbox project.