汉字拼形

https://shutonggui.cn/

https://hanziku.github.io/hzpx/

夺回数字空间的汉字主权 Seizing the Dominance of Hanzhi in Digital Space

汉字编码底层设计失误

拼音文字的单词(word) 是字母(alphabet)的一维组合（字符串）。汉字是部件[1]的二维组合。汉字和拼音词一样，都是开放集合(Open Set)。

但在目前计算机编码架构（Unicode）之下，汉字被误认为，和拼音字母一样的基本元素，即「码位」(Code-point)（码位可视为汉字的身份证号码，没有码位的字只能以图形，无法以文字形态立足於数字世界），这个错误的设计，导致一系列的问题：

1） 收字越多，码位就越多，字型就越大，键盘输入和手写输入的选取负担也越大。
2） 失去了像拼音文字随意造新词，简洁表达新事物的能力[2]。
3） 由於新字（在古文献数字化的过程中）不断地被发现和创造，每个新字都必须赋予一个单独的码位[3]，扩编永无止境[4]。
4） 每扩编一次，字型、输入法，甚至应用程序，都必须做相应的修改，否则看不到新字、也无法输入。一个新字从码位的赋定，到畅通无碍於数字世界，须等待好多年[5]。

过往尝试

「以一组数量较小（一千以下）的部件为构形单元，加上组合规则，产生无限的汉字。」

早在1973年，台湾谢清俊教授就提出了如上「字形产生器」构想[6]。西安张时钊先生在1984年也有类似的实作[7]。台湾的刹那工坊在2003年完成了黑体字形产生器，从Unicode 的IDS（表意文字描述序列）生成了「招财进宝」[8]。

美国的文林研究所设计的CDL（汉字描述语言）[9]，以XML为基础，可产生较为美观的字形。

---

[1] 组成汉字的部件。例「江」=氵+工。「氵、工」再往下拆解就无意义。组字可以多阶，例：鸿=江鸟。
[2] 例：Corona Virus Disease = Covid
[3] 有些汉字有非常多的异体字，如「宝」「龙」，Unicode的应对方式是，沿用标准字的码位，再後缀一个异体代码，这样就不必对每个异体字指定一个新的码位。详见 Unicode IVD。
[4] Unicode CJK (中日韩字符集) 2020年的第七次的扩编(Extension G)，加了4939个汉字。
[5] 截至2020年，绝大部份的手机都不支持 2009年的第三次扩编(Extension C)，甚至 2001 年发布的第二次扩编(Extension B)都不支持 (因为从Extension B之後，超过Basic Multilingual Plane 65536 个字符的范围，作业系统的文字显示层和应用程序须做相应之修改)。
[6] 「中文字根的貯存和中文字的合成」交大學刊 1973 二月。
[7] 无字库汉字系统 http://www.chancezoo.net
[8] 內嵌在Accelon3 系统中，從Unicode IDS式子「⿰辶⿱一⿱玨⿰隹⿰貝招」直接生成
[9] http://www.wenlininstitute.org Character Description Language

目前，日本上地宏一准教授[10] 开发的Kage 部件造字系统，虽然不是全自动[11]，但美观度和实用性较佳，基於该系统产生的花园字体，完全免费，是目前收字最齐全的字体。

汉字拼形

「汉字拼形」是输入「拼形式子」，产生「字形」的软件库。

与过去所有字形产生器的关键差别是，采用部件「置换」，而不是部件「叠加」来来描述汉字。比方说要组  ，叠加法列出部件及组合方式，例如「⿰艹⿰冏月」(「冏」「月」左右组合之後，再和「艹」上下组合，这是Unicode IDS语法）。

以「汉字拼形式」表达，「萌日冏」□  ，意思是将「萌」中的「日」，替换成「冏」。「萌」是此式子中的基础字，简称基字。

这解决了字形产生器的两大痛点：

1) 是部件输入不易，如「辶」、「宀」、「琺」，一般人不会念，即使会念，所熟悉的输入法也不一定支持。

2) 部件的比例分配及笔划变异的问题。比方说，「冏」和「月」组合时，「月」的撇笔要向左拉长。门与柬组合，门内的空间应扩大[12]。

由於「朙」是「明」的异体。理论上，从「明」衍生的字，如奅、盟都可以替换部件，衍生新字的数量远比想像的多，采用汉字拼形，皆可表达:「盟日冏」得  ，「奅日冏」得 。

替换的部件也可以是另一个拼形式子，利用这个特性可以组合复杂的合文。

例：「邏羅寶貝鼸從瞰致招」□  。也就是基字「邏」后面的字为一减一加，即「邏羅寶貝鼸從瞰致招」為「邏-羅+寶-貝+鼸−從+瞰-致+招」，意思是:將「邏」中的「羅」替換為「寶」，「寶」中的「貝」替換為「鼸」，「鼸」中的「從」替換為「瞰」，「瞰」中的「致」替換為「招」。有点繁琐，但对计算机来说只是很简单的套叠。

由於基字本身就隐含了部件的比例分配和布局信息，不必使用复杂的布局算法，也免去了人工微调，可立刻产生相当美观的字形。

在汉字拼形的安排之下，任何新字，就只是由系统已有字，所构成的字符串，人们可以像任意创造拼音单词一般，自由地创造及交流新汉字，不必等待Unicode 组织开会讨论扩编，这对汉字在数字世界的自主权和生命力有重大的意义。

实现方式

目前以Javascript实现的汉字拼字的程式及数据，约9MB，只能藉由Google Chrome浏览器以「外挂」的方式输入及显示，由於中国无PC作业系统之主导权，从通讯软件及手机平台是唯一可行的切入点:

[10] https://kamichi.jp Kage 系統
[11] Kage 系统还需人工选取部件，微调比例。
[12] 「柬」太小

1） 先置入微信，让大家自由造字，即时分享。
2） 手机浏览器原生支持，须修改UCBrowser 文字显示层。
3） 修改文本布局库（如Pango），并植入手机的图形界面（如鸿蒙）。
完成了汉字拼形的「内循环」之後，Google, Microsoft, Apple势必从善如流。

Q&A

问：为什麼要显示错字？

所谓的错字，只是字典中没有收录，或者相关部门规定不同的字，并不代表用不到此字。

在中文被视为的错字，可能在日文或韩文是正确的字。（例：日文「步」多了一点）

不如将错字与正字一视同仁，教科书编辑者可省去Photoshop 制图的工夫。例：初为初之误。

问：现代还有造新字的需要吗？

𬌗，口腔医学用字，Occlusion之译。1945年造，直到2015年才收进第五次扩编（Extension E）

砼，混凝土，即「人工石」，音「同」，1953年造，建筑业常用字。

2017年5月，国家语言文字委员会发布了四个新的化学元素用字。鉨鏌鿫鿬，原子序数分别为113， 115，117，118。

问：新字如何检索及输入？

「汉字拼形式」可展开成部件集，将「萌日囧」⿱萌囧，展开为「萌、艹、明、日、月、囧」，使用者可用熟悉的输入法，输入一个或多个部件即可检出。具体的实现有台湾开放古籍协会康熙字典的「部件」搜寻功能，「汉文博士」软件也有类似功能。

问：不受限制地创造新字，是否增加学习的负担？

在没有计算机的时代，创造新字是很自然的事，唯有经过时间汰选，构形合理优美、表达能力强的，成为现今常见、应学的字，而大部份的字只是昙花一现，然後静静地躺在字典里。

「汉字拼形」并不是标新立异的构想，它只是为了让汉字，回到和拼音文字一般，头等公民之地位。

以下饱含新字的句子，诸君能识否？

⿹气电車只是過渡方案，⿰车电最終將完全取代取代⿹气車。

⿸广⿰犭新改變了世界的格局，為⿰目英⿰目美⿰目加⿰目新⿰目奥所不樂見。

提案人：葉健欣 yapcheahshen@gmail.com

微信號： Sukhanika