最新一代中文输入法—汉字词语码编码技术

曾养志 曾巍 曾嵘

云南农业大学

【摘要】本技术根据汉字起源于象形文字，经指事、象形、形声、会意、转注和假借而发展演化、始终具有“形声相益”的特性，以汉语言文字中能独立存在和运用的单字、双字、三字及四字以上词组、词语、短语、诗词和科技词语等作为一个编码单位，采用“反切相拼定音节，声母加形识末字”的方法，使汉语因同音字、词太多而存在的大量重码得以识别，从而实现词语、成语、诗词以及科技词汇的盲打；词语编码改变后缀，可快速切换对应英语等外语词汇；用单位简称编码改变前缀或后缀，可切换全称或对应外语名称。本编码原理符合汉语言构词特性和发音逻辑思维，平均码长短，易学易用，得心应手，录入速度快。除普通版本外，可按不同学科专业延伸专业词语编码。本编码除适用于中文输入和信息处理及各类电子词典、辞书及汉英词典词条的检出外，也可开发为手写编码检出汉字词语，同时又是学生学习外语的强有力工具。编码熟悉后，还可作为新闻记者、文秘人员和学生的速记码。

电子计算机问世以来，汉字的计算机输入技术一直是中文信息处理的关键。大量常规信息——报纸刊物、书籍、科技文献、电子图书、日常公务、网络信息、商业贸易及工农业生产和民众生活等，无不涉及汉字的电子计算机处理。因此，汉字的编码成了中国汉语言文字工作者及计算机制造厂家重要的研究课题。八十年代初，王永明率先推出了“五笔字型输入法”，随后又有全拼、双拼、自然码、智能拼音等编码问世。迄今，已专利的汉字输入法达数百种。五笔字型、拼音码等较优秀的编码已得到广泛推广，为中国的计算机汉字输入和信息处理作出了重要贡献。然而，目前已推广应用的汉字编码技术，多数仍停留在汉字特有的“单字”处理上，还没有一种完善的、真正体现汉语特点并以词语为主的输入技术。对此，国家语委会主任许嘉璐曾指出：“目前，中文信息处理虽然已实现了‘字处理’，但这只是信息处理的初级阶段。要实现计算机中文信息的高速处理，....就必须超越‘字处理’阶段，对（汉语）语言中的词、短语和句子以及语言的应用规律进行深入的研究，同时，在此基础上制定电子计算机所用的语言和文字规范与标准”（《科技日报》1997年）。微软公司中国研究院院长李开福在向比尔.盖茨作述职报告时指出：“.....我们在寻求更好的计算机中文输入方面所做的工作。”他接着“着重介绍了中文输入方面的研究，谈到了不同的中文输入方法。我发现，比尔早已明白了中文输入的困难所在以及拼音和五笔等输入方法的利弊。我指出，如果中文输入的速度提高一倍，在每两小时的工作时间里，我们就可以帮助中国的计算机用户节省10亿个小时”。以上指出了汉字输入法的现状及希望通过研究所要达到的目的。

目前的汉字编码不外如下系统，即：区位码和电报码，由于不易记忆又只能录入单字，难以推广；形码，以五笔字型为代表，这是以构成汉字字型结构的笔划及所谓“字根”进行汉字拆分，再与键位和

“区”给定特定的码号与相应的键名对应。经过培训的专业人员具有较快的录入速度。五笔字型的发明在特定历史时期功不可没，其应用和普及程度也较广。然而，由于其编码着眼点为汉字的单字，不可能真正解决词语的编码。拼音码有全拼、双拼和智能拼音等。无论那种拼音码，其优点是不必拆分组合，拼读符合人们的听想思维习惯，编码反应直接，懂得拼音规则和韵母组合的代码键，上机即会，不用培训，不会忘记。然而，因汉语的发音仅限于418个音节，汉字字数太多，每个音节必然有大量同音字，当按下任一拼音组合的键位后，屏幕提示行即出现大量重码。尤其那些将单字、双字、三字乃至多字词组和短语都一律规定为4码者，当键入某一个编码时，大量的单字、双字词、三字词乃至多字词或短语就同时出现，录入者得反复地去“翻页”，有时须在多次“翻页”中的数百个词语里去寻找所需的那个唯一词组或词语。按《汉语拼音方案》设置的“全拼”码，韵母组合和词组拼写很多相混，如“xian”为“先、线、现、”等48个单字，而“西安、西岸、系按”也是同样的拼法。“xianshi”不知是“西安市”还是“现实、显示、现时、县市”。键入“ji”，出来“几、及、既、”等110多个单字，加一个“e”，则是“接、节、”等40多个单字，而录入者需要的是“饥饿”或“嫉恶如仇”等词组，用“jie”就拼不下去了。这种情形涉及整个汉语拼音音节中的很多部分。其次，汉语毕竟不是拼音文字，中国从小学学习汉语拼音，但很多大、中学生都不能掌握规范的拼音，用起来错误仍然很多。目前已推广的许多编码，由于存在大量重码，需要不断“翻页”，或因编码不科学、难记忆而影响录入速度，尤其是只着眼于单字的编码技术，是不能真正解决汉字的录入问题的。

形码是按汉字结构的基本笔画或繁杂部首进行拆分组合，拼音码虽然抓住了发声这一重要特征，然而，发声是任何一种语言都具有的特性。那么，汉字的特性是什么呢？中国的汉字，从新石器晚期的刻画符号开始，历经3000余年发展到殷商时期的象形文字，在象形文字基础上发展演化而成为系统的汉字。汉字发展演化的方法，就是所谓“六书”。“六书”者，即象形、指事、会意、形声、假借和转注。“象形者，画成其物，随体诘诎”。“指事者，视而可识，查而可见”。“会意者，比类合谊，以见指撝”。“假借者，本无其字，依声托事”。“转注者，建类一首，同意相受”（许慎：《说文解字》序）。及至现代，汉字虽然几经改革和简化，结构和数量发生了很大变化，但基本上仍保留上述特性。许慎在《说文解字》叙中说：“仓颉之初作书，盖依类象形，故谓之文。其后形声相益，即谓之字”。汉字虽非仓颉一人所能发明，但这段话却指明了汉字的形成是“依类象形”的。“形声相益”，就是汉字不仅具有“声”（发音）的特性，而且具有“形”的特性。汉语的发音为418个音节。所有的汉字，无论是8000余单字的新华字典，还是4万余的《康熙字典》，其发音都在这些音节范围。这就是汉字同音字-词多的根本所在。如此多的同音单字，如何区别每个字的意义呢？这就是依靠每个单字构成的“形”。这个“形”，一是由最早的原初字构成，其次是在原初字的基础上经指事、会意、形声等六书所衍生确立的偏旁部首。例如“丁”字是象形字，而现代语言中“ding”这一音节共有22个单字（《新华字典》），其中以“丁”字为发音基础加不同的偏旁部首，就构成了不同意义但仍然发“ding”音的单字就有18个，以“定”字加偏旁部首衍生的有5个。这些都发“ding”音的字如何区分呢？这就是汉字发明者赋予每个字的“形”——部首来加以识别。丁加口旁为叮咬的叮，丁加言旁为预订的订，丁加目旁为用眼睛盯上，丁加金属为铁钉的钉，丁加页（头）为顶，丁加田为町，丁加玉（王）为玉佩的响声玎，丁加病头为一种疮，丁加耳旁为耳垢的耵，丁加酉为酒泡的药剂酊，丁加水是一种水剂汀，丁加革为补鞋底的靪，丁加食旁是一种陈设的食品饤。部首一加，发音不变，意义却清楚明白，一目了然。以“登”（deng）字为基本字形加不同部首衍生出12个发“deng”音的单字，占这个音节的80%。“fang”这一音节共收单字19个，都是由“方”这一原初字加不同部首组合而成，这种情形在400多个音节中随处可见；另一种组合：则是由一个原初象形字作偏旁部首，再加不同的字组成，其发音随后面所加的那个字的读音而发，从而又构成了大量具有同类性质而意义不同的字的系列。如“牛”字是原初象形字，以“牛”作部首衍生的字，《说文解字》49字，《新华字典》收“牛”部为46字，其中许多字义已发生了变化。依此类推，凡人之属皆从人（亻、彳），凡草之属皆从艹，凡木之属皆从木等等，这就是汉字以部首分类的依据。所以，汉字只用一种特性如发音或“形”都无法反映单字和词组的特性。可以说，“形声相益”是汉字演化和扩展的主要方法，电子计算机时代的汉字编码也应以此作为基本的识别方法和原则。

现代汉语中，描述各类事物并以文字作为信息载体传递的文章，是由词组和短语构成的。实际上，国家标准局公布的6763个单字中，有1000余个是不能单用的，只是组词的单元，如“琵琶、枇杷、菠萝、荸荠”等。有些虽可单用，但组词后就很少单用了，这类单字也有1000多。因此，老是花功夫去研究单字是没有必要的。随机统计了有代表性的文稿，双字词占41.74%（35.8—46.9%），三字词占20.17%，四字词占21.76%，五字以上及短语占7.08%。单字仅占9.25%，包括虚词“的、地、和、与、及”等。此外就是科技论文和著作，这类文章除普通词汇外，有大量专业技术词汇和术语。随机统计了科技论著中的12397个字词，有双字词4636个，占总字数的37.39%，比普通文章略低。三字词2934个，占23.67%，比普通文章多。四字词1958个，占16.04%，大大高于普通文章中的5.25%。五字至八字词或短语749个，占6.04%，普通文章中仅占1.51%。这表明，科技文章中多字词语的使用频率比普通文章高，因科技词语本身就是以多字词为主的。国家编订的《汉语主题词表》（“自然科学”版），其中“B”这一声母16个音节共收入主题词3456个，其中双字词505个，三字词822个，四字词1017个，五字词567个，六字词290个，七字以上254个，单字却没有。科技文章中普通词汇与科技词汇之比为4624：6249=1:1.35。这表明，对于科技工作者，既使普通汉语词语的编码问题完全解决了，但在写作科技文章时，仍有一半以上的专业技术词汇须一个单字一个单字地录入。此外，各类电子词典，其汉字词语和汉英词典词条的检出，都只能一个个单字录入在显示屏上组合成词语后才能检出和汉英翻译。可见，汉字的编码如果忽视了科学技术词语的编制和研究，仍然是一种不完全的编码技术。不难看出，迄今推广应用的各种输入法，还没有真正解决汉语以词语为主的编码，更谈不上科学技术词语的编码了。

本技术根据汉字“形声相益”的特性和现代汉语词语的应用范围，以汉语中双字词、三字、四字以上词语、短语及固定的简单句作一个编码单位，采用“反切相拼定音节，声母加形识末字”的方法。“反切拼音”是我国宋朝以后用于汉字注音的方法，为一字之声母与另一字之韵母快速相拼，优点是简洁明快、节省码长，与“双拼”相似。“声母加形识末字”：双字词、三字词第一个字由“反切相拼”定其所在之音节，末字用其声母加部首识别。4字以上则只需反切相拼定音节，以后各字用声母组合。短语或中间有停顿的固定短句在停顿处用后缀省略。由于许多单字具有词的性质或有时可能单独用到，仍将单字编码列出。文章中最常用到的虚词、连词、形容词和付词用一键输入。编码方案如下：

（1）汉字偏旁部首的调整和“0”部首的设置：本发明采用“反切相拼定音节，声母加形识末字”的方法，因为末字需要用部首识别，而传统习用的偏旁部首多者为213部，少者也有188部。其中50余部不规范，不仅识别困难，且计算机的键位也难以合理安排，所以对传统部首进行了调整改进。即将各类字典中列为“难检字”表中的500余单字绝大部分划为“零”部首，用键名“o”键代表，有些则归入相应的规范部首。

（2）单字编码：词语码录入已很少用到单字了。但一些具有词汇性质的名词、动词及姓氏、名号和古汉语等涉及的单字仍不少，因此仍将其编码列出。单字用3码，编码规则是：“反切相拼定音节，重码部首来识别”，例如“中”字，汉语拼音为“zhong”，“双拼”为“vs”，当键入“zhong”或“vs”时，屏幕提示行出现“中、重、种、钟、肿、众、终、盅、忠、衷、踵、舯、螽、冢、塚”等同音字。本发明规则：反切相拼定音节为“vs”，若需其中某字时，只需在“vs”后面加该字的部首即可检出。“中”字部首不规范加“o”为“vso”，“重”字横底加“/”为“vs/”，“种”字禾旁加“h”为“vsh”，“钟”字金旁加“j”为“vsj”，“肿”字月旁加“y”为“vsy”，“众”字人旁加“r”为“vsr”，“终”字丝旁加“s”为“vss”，“盅”字皿底加“m”为“vsm”，“忠”字心底加“x”为“vsx”，“衷”字点上加“’”为“vs’”，“踵”字足旁加“z”为“vsz”，“舯”字舟旁加“v”为“vsv”，“螽”字虫底加“i”为“vsi”，“冢”字豕底加“u”为“vsu”，“塚”字提土加“t”为“vst”等。这样，上述15个单字的重码就可用各字所具有的部首区分而没有重码。

（3）双字词的编码：双字词用4码，编码规则是：“反切相拼定音节，声母加形识末字”。反切相拼定准第一个字的音节，第二字用声母加部首识别。如“实际”一词，先键入“ui”这一音节，接着键入“际”字的声母“j”，“际”字的部首为“耳”旁，“耳”的声母发音为“e”，所以加“e”为

“uije”。与“实际”发音相同者还有：“世纪”的“纪”为丝旁，加“s”为“uijs”，“时机”的“机”为木旁，加“m”为“uijm”，“事迹”的“迹”为走之，加“z”为“uijz”，“试剂”的“剂”为利刀，加“d”为“uijd”，“史记”的“记”为言旁，加“y”为“uijy”，“史籍”的“籍”字为竹头，加“v”为“uijv”，“诗集”的“集”为“佳”上，加“j”为“uijj”，“石鸡”的“鸡”字为鸟旁，加“n”为“uijn”，“实据”的“据”为提手，加“t”为“uijt”，“诗句”的“句”为口部，加“k”为“uijk，“时局”的“局”为尸头，加“u”为“uiju”。这样，发音相同必然为重码的词组就用部首切分开了。双字词中尚有少量同音词组的末字同形，也会出现重码，但未超过6个，经再次拆分即可做到全无重码。

（4）三字词的编码：三字词用5码，其规则是：“反切相拼定音节，声-声加形识末字”。词组第一字反切相拼定准其音节，第二字用声母，第三字用声母加部首识别。例如：“中国人”为“vsgro”，人字部首不规范为“o”），“中国热”为“vsgrh”，“中国字”为“vsgzg”，“中国话”为“vsghy”，“中国画”为“vsgh/”；“第一班”为“diybw”，“第一版”为“diybp”，“第一榜”为“diybm”等等。

（5）四字及四字以上词组或短语的编码：四字词编码用5码，五字词用6码，六字词用7码，七字词用8码，八字词用9码，九字以上用10码。其规则是：“反切相拼定音节，其余声母来识别”。第一个字反切相拼定准音节，其余各字用其声母识别即可，例如：“中国人民”为“vsgrm”，“人民解放军”为“rfmjfj”，“百闻不如一见”为“blwbryj”，“不到长城非好汉”为“budiifhh”，“矮子里面拔将军”为“alzlmbjj”等等。

（6）固定短语和句子的编码：固定短语和句子多在成语或名人名言中出现，如“横眉冷对千夫指，俯首甘为孺子牛”等。编码规则：前半句编码加后缀。如“横眉冷对千夫指，俯首甘为孺子牛”为“hgmldqfvv”。如只需前半句，则前半句编码键入后不重复最后那个编码即可。有两次停段者，全句再加一相同码号，如“不破不立，不塞不流，不止不行”为“bupbll”。只需“不破不立，不塞不流”为“bupbll”。

（7）诗词曲赋的编码：从“诗经”、“楚辞”开始的诗词曲赋，是我国文化艺术的瑰宝，历来为人民所喜爱，常用于学习、咏颂和引用。编码规则：以完整诗句编码上半句，需全句则加后缀。如“红军不怕远征难”为“hsjbpyvn”，若要整句“红军不怕远征难，万水千山只等闲”则为“hsjbpyvnn”。五律、七律要全诗一次出现，只需在第一半句编码后加/q，如moscmkjs/q即为“暮色苍茫看劲松，乱云飞渡仍从容。天生一个仙人洞，无限风光在险峰”。对于词赋，则按自然语句录入。

（8）单位及机构名称编码：在新闻稿件及公务文件中，常涉及世界各国、国际组织、政府机构、高等院校、研究院所、金融财贸和工商企业等机构名称。这些名称一般常用简称，如发表公报、签订条约或协议等又须用全称。“美国”为简称，“美利坚合众国”为全称。“法国”为简称，“法兰西共和国”为全称。“中共中央”为简称，“中国共产党中央委员会”为全称。“全国人大”为简称，“全国人民代表大会”为全称等。编码方案：一律用简称编码，需全称时加后缀/q。“美国”为“mzgo”，“mzgo/q”则为“美利坚合众国”。“波黑”为“bohh”，“bohh/q”则为“波斯尼亚和黑塞哥维那共和国”。“中共中央”为“vsgvy”，“vsgvy/q”则为“中国共产党中央委员会”。上述国家、国际组织、政府、高校、研究院所以及单位、部门的简称，必须用公知公用的简称，否则不能正确检出。

（9）词语和机构名称切换英语等和科技拉丁语的编码：科技文化、经济贸易、旅游、新闻媒体和各行各业都涉及外语，记者也常在文章中直接用英语陈述，尤其是各类电子词典中汉英词典词条的检出等，都需要按中文词语原意译成英文。本编码发明了汉语词语、科技词语和机构名称的英语、法语和科技拉丁语快速切换方法。编码为：在汉字词语和机构名称简称编码基础上，加后缀或改变后缀即可。英语加后缀/e，拉丁语加/l，法语加/f，德语加/g，西班牙语加/s等等。先按编码规则输入汉语词语，屏幕出现该词语的汉字词条，如只需该汉字词语，则击空格键上屏，若需将该词语转换为英语，则不击空格键而在其编码后加后缀/e，即出现相应的英语。如：mzgo—美国，mzgo/q—美利坚合众国，mzgo/e—the uited sates，mzgo/eq—the uited sates of america。yngo—英国，yngo/e—britain，yngo/q—不列颠和北

爱尔兰联合王国，yngo/qe—united kingom of great britian and northenr ireland等等。科技文章需要英语、拉丁语学名索引或注释时，只需在该词语的汉字编码后改变后缀即可。

（10）外语缩写作编码：报章、教材和科普文章中，常用英语缩写如WTO，FAO，DNA，RNA，APEC，CCTV等。编码方法：用缩写原文加前缀和后缀构成。英语加e，拉丁语加l，后缀则根据需要而变化。如efao—粮农组织，efao/q—联合国粮食和农业组织，efao/e—food and agricultural organization of United Nations 。eapec—亚太经合组织，eapec/q—亚洲和太平洋地区经济合作组织，eapec/e—Asia and Pacific ocean area economic cooperte organization。ecctv—中国中央电视台，ecctv/e—China centre televition 。eopec—石油输出国组织，eopec/e—the organnization of petroleum euporting countries等等。

本编码容词量大，编码字典第一版收词语13万余，双字词组达47000余，重码最多6个。

井田汉字，独一无二的汉字结体构形理论，能够科学地解决数码时代汉字所面临的问题！