语言服务书系 · 华文与华语教育

东南亚华文媒体用字用语研究

刘华 著

语言服务就是利用语言（包括文字）、语言知识、语言艺术、语言技术、语言标准、语言数据、语言产品等等所有语言的所有衍生品，来满足政府、社会及家庭、个人的需求。 ——李宇明

暨南大學出版社
JINAN UNIVERSITY PRESS

语言服务书系 · 华文与华语教育

东南亚华文媒体用字用语研究

刘华 著

暨南大學出版社
JINAN UNIVERSITY PRESS
中国 · 广州

图书在版编目（CIP）数据

东南亚华文媒体用字用语研究/刘华著. —广州：暨南大学出版社，2015.9
ISBN 978-7-5668-1328-2

Ⅰ.①东… Ⅱ.①刘… Ⅲ.①汉语—语言学—研究—东南亚 Ⅳ.①H1

中国版本图书馆 CIP 数据核字（2015）第 019136 号

东南亚华文媒体用字用语研究
著　　者　刘　华

出 版 人　徐义雄
策划编辑　杜小陆　刘　晶
责任编辑　杜小陆
责任校对　黄　斯
出版发行　暨南大学出版社（广州暨南大学　邮编：510630）
网　　址　http：//www. jnupress. com　http：//press. jnu. edu. cn
电　　话　总编室（8620）85221601
　　营销部（8620）85225284　85228291　85228292（邮购）
排　　版　广州良弓广告有限公司
印　　刷　佛山市浩文彩色印刷有限公司
开　　本　787mm×960mm　1/16
印　　张　14
字　　数　273 千
版　　次　2015 年 9 月第 1 版
印　　次　2015 年 9 月第 1 次
定　　价　35.80 元

前　言

国家语言资源监测与研究中心“海外华语研究中心”（教育部语言文字信息管理司与暨南大学共建）自2005年成立以来，一直致力于构建“海外华语语料库”。在此语料库基础上，中心进行了一系列海外华语方面的研究，本书就是成果之一。

首先要感谢教育部语言文字信息管理司、国家语言资源监测与研究中心的大力支持。李宇明教授、王铁琨教授一直关注“海外华语研究中心”的发展和成长，对作者个人也是关怀备至，感谢他们对中心、对本人学术成长的指导和支持。

感谢商务印书馆的魏励老师、蔡长虹博士、刘建梅博士，他们对本书的内容提出了很多建设性的建议。

感谢国家语言资源监测与研究中心平面媒体语言分中心的张普教授与杨尔弘教授、国家语言资源监测与研究中心有声媒体语言分中心的侯敏教授、国家语言资源监测与研究中心教育教材语言分中心的苏新春教授与郑泽之教授、国家语言资源监测与研究中心网络媒体分中心的何婷婷教授、国家语言资源监测与研究中心少数民族语言分中心的赵小兵教授，和他们一起研究讨论的日子里，他们教给我很多知识，让我这个晚辈收获良多。

感谢暨南大学华文学院院长、“海外华语研究中心”主任郭熙教授，感谢他为我们这些年轻人搭建了一个很好的学术平台，感谢他在生活上、学术上对我的关心和支持。同时，感谢“暨南大学华文教育研究院”为本研究提供经费资助。

最后，我要特别感谢一直以来默默关爱和支持我的亲人和朋友，是你们无私的爱陪我走到今天，谢谢你们！

本书基于大规模语料库，利用计量方法，对海外华语字词进行统计研究，希望能为读者提供语料库计量研究方法和海外华语字词使用数据方面的参考。

本书部分成果曾发表于国家语言资源监测与研究中心编的《中国语言生活状况报告2008》（下编）（商务印书馆，2009年）、教育部语言文字信息管理司组编的《中国语言生活状况报告2011》（光盘版）（商务印书馆，2011年）。本次出版对其重新做了修订。

语料的例句检索和字词检索参见网址：http：//www. globalhuayu. com。
欢迎大家提出宝贵意见。

刘　华
于暨南大学华文学院
2014 年 10 月 10 日

目　录

第一章　绪　论

本章主要回顾了华语、海外华语及海外华语传播的研究现状，综述了海外华语的研究成果，特别总结了海外华语在词汇方面的研究。

第一节　华语与华语传播

一、华语、海外华语

关于“华语”一词的定义，历来争议颇多。20 世纪 80 年代以来，陆续出现了“华人的共同语”（陈重瑜，1986）、“汉语在海外的通称”（田惠刚，1994）、“全世界华人的共同语”（周有光，1995）等多个内涵和外延各不相同的定义。关于“华语”的定义问题，郭熙（2004）已有较全面、充分的论述，本书不再赘述。本书中的“华语”采用郭熙的定义：华语是以现代汉语普通话为标准和核心的华人共同语（郭熙，2006）。同时，本书所研究的“海外华语”的应用范围限于海外各个国家的华人社会，不包括港澳台地区。

二、海外华语传播与现状

20 世纪 80 年代以来，随着中国综合国力和国际地位的不断提升，国际上兴起了一股持续性的“汉语热”。据有关部门公布的数据，截至 2005 年，世界上通过各种方式学习汉语的人数超过 3 000 万，而这一数字仍在不断上升。其中，绝大多数的学习者是华人。[①] 在积极开展对外汉语教学、进行汉语国际推广的同时，我们认为应当对海外华语的生存状况进行系统性研究，以推动汉语的传播、文化的传承以及相关部门政策的制定。

海外华语的使用环境相对于大陆而言要复杂得多。一方面，海外华语始终与闽、粤以及客家等汉语方言共存，如很多海外华语学校使用粤方言等方言进行教学，而且在华语使用过程中，繁简字并用、注音不规范等现象长期存在；另一方

① “中国语言生活状况报告”课题组．中国语言生活状况报告 2005（上编）［M］．北京：商务印书馆，2006.

面，海外华语处于多语环境的包围之中，与其他语言的互相影响和融合在所难免，因而在语音、词汇、语法等多个方面都呈现出与标准的汉语共同语不同的面貌。[①] 例如，在语音上，声、韵、调都与中国大陆的标准普通话有所不同，有入声而无轻声和儿化；语法上，存在类似于闽方言和粤方言的“V + Adv”、“有 + V”、“V + O + 一下”等句式；词汇上，除一些表达海外华人社会特有概念的词语外，还有许多词语与普通话名称相异而意义相同或相近，例如，马来西亚华语中的“卫生所”指殡仪馆、“饭盒”指盒饭，新加坡华语中“两造”指双方、“灵犬”指警犬等。[②] 其中，尤以词汇方面的分歧最大，借词的现象最为突出。这些方面的差异不仅给华语学习者在学习过程中造成了许多困难和障碍，同时在海外华语和汉语标准语之间形成了理解上的差异和分歧。这些差异和分歧不仅仅体现在汉语学习和教学上，在更深层次的意义上这些差异和不协调也直接影响到汉语的国际推广和我国相关部门语言政策的制定。因此这个问题值得我们重视和认真研究。

第二节　海外华语研究现状

东南亚华语及华语文教育的研究尚处于起步阶段。国内主要集中在暨南大学、华侨大学和海外华语研究中心（教育部语言文字信息管理司与暨南大学共建，2005 年成立），海外主要以华语桥为基地，聚集了一批华语及华语文教育研究的学者。

目前，东南亚华语的研究主要集中在以下几个方面：华语的界定、性质研究（张从兴，2003；郭熙，2004、2006；陆俭明，2005），华语语言特点研究（陆俭明，1996；周清海，2000；徐杰，2004），华语区域词语、特色词语及变异研究（周清海，2002；曾晓舸，2004；汤志祥，2005；刘文辉，2006），华语和现代汉语对比研究（周烈婷，1999；邢福义，2005；贾益民，2005），华语规划与华语规范研究（谢世涯，2000；林万菁，2001；郭熙，2002、2006），华语推广与华语文教学研究（郭熙，2007）。其中华语研究的地域来源主要是新加坡、马来西亚、泰国和印度尼西亚等地。

字词是语言研究的基础，字词表更是语言教学的根基。东南亚华语字词的研究主要集中在“社区词”、“词源与词语对比”和“字词使用规范”三大块，如港澳社区词研究、新加坡社区词研究、词语探源、华语与汉语的词语对比、华语

① 郭熙．华文教学概论［M］．北京：商务印书馆，2007.

② 郭熙．域内外汉语协调问题刍议［J］．语言文字应用，2002（3）．

的规范与协调等等。其中，由李宇明主编，众多海内外华语研究学者联合编撰的《全球华语词典》是其中的代表作。

针对海外华语同现代汉语标准语的分歧和差异问题，目前已有一些研究，例如陈松岑（1996）、李如龙（1996）、邹嘉彦（1996）、陆俭明（1996）、郭熙（2000）、汪惠迪（1999）等学者的研究。但总体来说，研究还不够全面、深入，缺乏系统性。主要问题在于缺乏关于海外华语分布和特点的第一手详细调查资料，对于现状的描写多偏重理论分析，所用语料大多来自作者自身的体验和总结，多从经验出发，比较单薄。由于缺乏大规模语料的支撑，对于海外华语的描述和研究还远不够充分与科学。

总的来说，东南亚华语词语的研究集中于对个别字词的探源，或是对某个海外社区的字词描写，或是华语独有词语的研究，尚未见到概括整个东南亚华语字词的研究。另外，在方法上大多是卡片式、个案式的专家经验式研究，尚未进行基于大规模真实语料库的计量研究。

面对这种研究窘境，建设一个大规模的语料库以满足研究需求就显得迫在眉睫。由于口语语料的收集和转写较困难，基于目前的研究条件和手段，书面语料是一种较合适的研究对象。而最能鲜活、动态地反映语言面貌的莫过于媒体语料。同时，媒体语料也能在一定程度上反映书面语和口语两种语体的面貌。因此，研究海外华语可以从研究海外华语媒体语料入手。在新媒体日益兴起的今天，网络成为我们最容易接触到、最具活力和影响力的媒体。网络媒体较传统媒体而言，信息量更大，语料更易获取，同时网络媒体的互动性确保了使用者语言面貌的真实性。因此，我们选择海外华语网络媒体（含报纸网络版）语料作为本书的研究对象。

第二章　语料介绍、方法说明与术语说明

本章具体介绍东南亚华语语料库的总体情况，对全书涉及字词描写的术语进行说明。

第一节　语料介绍

为了跟踪研究海外华语的使用情况，海外华语研究中心从2005年开始建设海外华语语料库。2009年，海外华语研究中心对东南亚华语语料库进行了用字用语的调查研究。

东南亚华文媒体较多，由于我们在语料获取上受到技术限制，加上其他因素的影响，有的华文媒体的语料无法获得。本次媒体的选择主要考虑了语料的可获取性、媒体影响程度和信息量三个因素。

本次调查的语料仅限于较有代表性的新加坡、马来西亚、泰国的主要华文媒体的语料（下文统称为“华文语料”）。语料时间跨度为2005年到2008年，均来自于网络，我们对其做了去除HTML标签信息和广告信息的处理，抽取了网页正文、标题、发表时间等信息。总文本数[①]为296 355。

下面是语料的具体信息（括号里为文本数）：

新加坡：亚洲新闻网（61 197）、新动网（26 228）、《联合早报》（63 697）；

马来西亚：马新社中文网（29 964）、《光华日报》电子新闻（63 346）、独立新闻在线（8 474）；

泰国：《世界日报》（43 449）。

为了更好地研究华语的特点，我们同时进行了华文语料与中国国家语言资源监测语料库语料（下文统称为“监测语料”）的比较调查。监测语料来自国家语言资源监测与研究中心平面媒体语言分中心和网络媒体分中心2005年到2008年的语料[②]，共4 474 675个文本文件，3 709 908 405字次[③]（不含部

① 所有文本文档的数量。

② 详细情况请参看：国家语言资源监测与研究中心．中国语言生活状况报告2008（下编）［M］．北京：商务印书馆，2009.

③ 调查语料中汉字出现的次数。

件），2 145 386 164词次[①]。

第二节　方法说明

本次的调查对象包括华文语料的汉字和词语，调查时以中国大陆汉语字词使用规范为参照。调查项目主要包括频次、频率、文本数、使用率、覆盖率等，并和监测语料进行了共用、独用、频序比的对比分析，还将华文语料的汉字统计结果和《现代汉语常用字表》、《现代汉语通用字表》进行了比较分析。

同时，进行了华文语料特色词语的调查研究，形成了《华文语料特色词语表》，并列举词语的提示性释义、例句、频次和出现文本数，以及进行了华文语料字母词的调查，形成了《华文语料字母词表》。

第三节　术语说明[②]

一、频次、频率、文本数

1. 频次

频次指的是调查对象在调查语料中出现的次数。如在华文语料中，汉字“的”总共出现了5 028 063次，其频次即为5 028 063。频次是语料库语言学中描写字词统计量最基本的参数，也是其他统计量，如频率、覆盖率等计算的基础。

2. 频率

频率指的是某一调查对象的频次与整个语料所含调查对象总频次的比值。如在华文语料中，所有汉字的总频次为161 728 981，汉字“的”的频次为5 028 063，其频率即为5 028 063/161 728 981 = 0.03 11（3.11%）。频率反映的是字词在语料中的基本分布情况。

3. 文本数

文本数指调查语料中某一调查对象出现的文本或文档的个数。如在华文语料

① 调查语料中词语出现的次数。

② 本节主要参考了《语言资源监测与研究相关术语》，见：国家语言资源监测与研究中心．中国语言生活状况报告2009（下编）［M］．北京：商务印书馆，2010.

中，汉字“的”总共在278 204个文本文件中出现过，其文本数即为278 204。文本数是对频次的补充，是反映字词使用范围，即文本分布的重要参数。有时候，频次较高的字词，如果其文本数较少，则说明其在文本中出现得相对集中，其真实的使用率相对低一些。

二、累加频率、覆盖率、使用率

1. 累加频率

累加频率指的是调查对象按频率排列，依次相加所得到的值。频率一般按降序排列。如某统计中将汉字按频率降序排列，前三位分别为：“的”，频率3%；“是”，频率2.9%；“大”，频率2.7%，那么，截止到“大”字的累加频率即为3% +2.9% +2.7% =8.6%。

2. 覆盖率

覆盖率指的是调查语料内指定调查对象数量占所有调查对象总量的百分比。如《中国语言生活状况报告2005》（下编）将汉语常用词语按照频次降序排列，前4 179条词语占了总调查语料9亿字的80%，那么这前4 179条词语的覆盖率就是80%。

3. 使用率

使用率指的是某一调查对象分布率和使用频率的综合计算值。使用率越高，分布越均匀，使用率与频次也就越接近。否则反之。计算公式如下：

$$D_i = t_i / T;\ U_i = F_i \times D_i$$

其中，D_i 是 i 号字的分布率，t_i 为 i 号字的出现文本数，T 为所有语料的文本总数；U_i 为 i 号字的使用率，F_i 为 i 号字的频率。

为了使得所有字的使用率总数为1，进行了归一化：

$$U_i = F_i \times D_i / \sum_{j \in V} (F_j \times D_j)$$

其中，F_i 为 i 号字的频次，分母为归一化项，V 表示所有字种。

三、频序、频序比

1. 频序

频序指的是某一调查对象在不同语料中按频次、频级或频差排列的顺序。如“大”字，在华文语料按频次由高到低排列的字表中，顺序为3，则其频序即为

3。本书中，频序指的是按频次排出的顺序。

2. 频序比

频序比指的是某一调查对象在不同语料中按频次排列的位序的比值。即将所有调查对象按频次从高到低排列，用调查表中某调查对象的位序值除以参照表中相同调查对象的位序值，得到的就是该调查对象的“频序比值”，即“频序比”。

在进行华文语料和监测语料的对比研究时，将考察范围内的汉字的频序比从低到高排列，可以得到华文语料中出现频序相对于监测语料相差较大的汉字，这在一定程度上反映了华文语料用字的特点。

例如，“坡”字在华文语料按频率由高到低排列的字表中，顺序为223，其频序即为223；在监测语料中，“坡”字频序为1 416。因此，“坡”的频序比值即为223/1 416 =0. 16。将华文语料和监测语料中共同使用的汉字进行频序比值的计算，最后将计算结果从低到高排列，前100个汉字如下：

党　坡　政　马　湾　国　府　选　台　吉　扁　泰　席　拉　美　民　及
议　令　表　阿　示　统　指　亚　他　伊　巴　隆　朝　警　陆　督　岸
贪　宪　印　吁　阵　官　早　哈　总　鲜　日　阁　须　加　署　玛　港
媒　曼　举　威　论　禽　说　华　恐　陈　怖　票　会　捕　将　独　长
谈　括　联　透　反　述　否　言　必　宗　洲　宣　立　尼　局　沙　谷
希　讨　达　盟　兹　该　贸　因　至　军　露　炸　报　若　新

四、字种、字种数、词种、词种数

1. 字种

调查语料中不重复的汉字。在中文信息处理中，相同字形的一般计算为一个字种。如“长短”的“长”和“首长”的“长”为一个字种。

2. 字种数

调查语料中不重复的汉字个数。如《中国语言生活状况报告2005》（下编）调查的所有语料中字种数为8 128个。

3. 词种

调查语料中不重复的词。在中文信息处理中，目前仍暂按词的书写形式来区

分词语，即相同词形的一般计算为一个词种。如表“进入水中”义的“下水”和表“食用的动物内脏”义的“下水”在统计中为一个词种。

4. 词种数

调查范围内不重复的词语个数。如《中国语言生活状况报告 2005》（下编）的用字用词调查中词种数为 1 651 749 个。

五、共用、独用

1. 共用

某一调查对象在全部调查范围内皆有使用。如当覆盖率达到 90% 时，华文语料中的异体字共 45 个，监测语料中的异体字共 39 个，二者共用的异体字为 12 个。

2. 独用

某一调查对象只在某一调查范围中使用。如当覆盖率达到 90% 时，华文语料中的异体字共 45 个，监测语料中的异体字共 39 个，华文语料独用的异体字共 33 个。

第三章　华文语料用字调查研究

本章对华文语料中的汉字使用情况进行了描述，并对按频率和按使用率排序所得字表进行了比较，同时对华文语料汉字的覆盖率和字种数关系进行了分析。

本次统计没有甄别文本中的别字、乱码，以及无法显示的字符，也未区分多音字、同音字。

华文语料中所有字符的总次数为 213 961 939，字符种数为 9 652。其中汉字总频次为 161 728 981，汉字字种数为 8 429（不含汉字部件）。

监测语料共 4 474 675 个文本文件，汉字总频次为 3 709 908 405（不含汉字部件），字种数为 11 802。

第一节　汉字使用的分类情况

作为语料来源的新加坡、马来西亚和泰国的汉字使用标准基本与中国大陆相同。从中国大陆汉字使用的视角来观察，汉字使用主要包括规范字和非规范字的使用。

规范字是指经过整理简化并由国家以字表形式正式公布的简化字和未被整理简化的传承字；非规范字指的是规范字以外的汉字。本调查中，非规范字以繁体字、异体字为主，也包括其他类型的非规范字，如旧印刷字形、日本汉字、旧计量用字、韩国汉字等等（由于其他类型的非规范字的字种数和频次都很低，因此本书中将之合计为“其他字”一类）。

目前学界对于汉字使用分类的研究尚无定论，特别是对于繁体字、异体字等争论较大。由于我们需要对华文语料和监测语料进行汉字使用分类的平行对比研究，监测语料字表中的汉字已按上文的汉字使用分类方法进行了分类，因此，对于华文语料，我们也采用监测语料字表中的汉字分类方法对汉字进行了分类，本章中并不深究汉字使用分类的学理上的根据。

一、概况

华文语料中，规范字频次为 161 692 898，字种数为 7 173，规范字频次在汉