基于大规模语料库的《古籍汉字分级字表》研究*

许长伟
南京师范大学文学院
changweixu36@gmail.com

冯敏萱
南京师范大学文学院
fennel_2006@163.com

李斌
南京师范大学文学院
gothere@126.com

袁义国
南京师范大学文学院
lexcliff1023@gmail.com

摘要

《古籍汉字分级字表》是基于大规模古籍文本语料库、为辅助学习者古籍文献阅读而研制的分级字表。该字表填补了古籍字表研究成果的空缺，依据各汉字学习优先级别的不同，实现了古籍汉字的等级划分，目前收录一级字 105 个，二级字 340 个，三级字 555 个。本文介绍了该字表研制的主要依据和基本步骤，并将其与传统识字教材“三百千”及《现代汉语常用字表》进行比较，验证了其收字的合理性。该字表有助于学习者优先掌握古籍文本常用字，提升古籍阅读能力，从而促进中华优秀传统文化的继承与发展。

关键词：四库全书；古籍汉字；字表；汉字分级

The Formulation of *The graded Chinese character list of ancient books* Based on Large-scale Corpus

Abstract

The graded Chinese character list of ancient books is a graded list based on a large-scale corpus of ancient books, which is designed to promote learners' ability to read ancient books. It offsets the vacancy of the research achievements of the ancient Chinese character list, which contains 105 first-class characters, 340 second-class characters and 555 third-class characters. This paper introduces the main theories and basic development steps of this list, and makes a comparison with the traditional words books and *The modern Chinese common character list* to verify the rotationality of the words contained. This list is helpful for learners to master the basic characters in ancient books and improve their reading ability, promoting the inheritance and development of Chinese excellent traditional culture.

Keywords: Siku Quanshu, Ancient Chinese characters, Character lists, Character lists

* 基金项目：江苏省社会科学基金项目（20JYB004）；南京师范大学教改项目新文科建设专项（2020NSDJG048）和重中之重项目（2021NSDJG002）

1 引言

字（词）表是指为专门目的而研制的特定字（词）集合，对学习用字（词）的掌握具有重要意义。最早的词表研制始于 20 世纪 30 年代美国，研究者将那些几乎在所有阅读材料中都出现的词，汇集成核心词表优先教学，使儿童的阅读更容易（Jerry，1970）。

我国历来就有确定常用字表的传统，孙钧锡（1991）在《中国汉字学史》中指出：“过去各个朝代编撰或流行的识字书，可以认为是当时的‘常用字表’。”作为我国古代最负盛名的蒙学教材，《三字经》、《百家姓》、《千字文》（简称“三百千”）选取了当时的常用汉字供儿童集中学习（陈黎明，2010），具有重要的价值,但这些识字教材通常由专人或专门机构编写，汉字的选取主要依靠个人经验或集体智慧，而非基于大规模的语料统计，存在一定的主观性和局限性。

蒙学教材	字型数	字例数
《三字经》	532	1140
《百家姓》	501	568
《千字文》	1000	1000

表 1.“三百千”收字情况

采用科学统计方法进行的常用字研究始于二十世纪 20 年代（费锦昌，1988）。1921 年，著名教育家陈鹤琴统计了白话文中汉字出现的频率，并于 1928 出版《语体文应用字汇》，这是我国第一本现代汉字字频统计专著，为汉字的计量研究做出了宝贵贡献（苏培成，2001）。

1988 年，国家语委完成《现代汉语常用字表》和《现代汉语通用字表》的制定，这是当代常用字表研制的重要里程碑。《现代汉语常用字表》共收字 3500 个，分 2500 个常用字和 1000 个次常用字。《现代汉语通用字表》共收字 7000 个，包括 3500 个常用字。

除常用字表外，面向不同应用领域的专业字表也层出不穷，且兼顾了分级需要。

字表	分级情况				
《汉语水平汉字等级大纲》	等级	甲级	乙级	丙级	丁级
	汉字个数	800	804	590	670
《汉字应用水平测试字表》	等级	甲级	乙级	丙级	
	字序号	1-4000	1-4500	1-5000	
《汉语国际教育用音节汉字词汇等级划分》	等级	一级字	二级字	三级字	
	字序号	1-900	901-1800	1801-2700	

表 2.其它专业字表及分级情况

然而，上述字表均基于现代汉语语料研制，用于对外汉语教学，汉字应用水平测试等目的，针对古籍文献进行的字表研究及成果则较少。

党的十八大以来，中华优秀传统文化的继承和弘扬被提到了前所未有的高度，古籍文献作为中华优秀传统文化的重要组成部分，其阅读和推广具有重要意义。由于原始古籍文本均用繁体字排版，接受简体字训练的现代读者，往往会出现阅读困难。本文基于大规模古籍文本语料库，统计构建了《古籍汉字常用字表》，并挖掘汉字分级计量特征综合分析，对其中的汉字进行分级，实现了《古籍汉字分级字表》的研制，该字表可以帮助人们优先掌握古籍阅读常用字，降低古籍阅读难度，从而促进中华优秀传统文化的继承与发展。

2 《古籍汉字常用字表》的研制

2.1 选字来源

字（词）表的研制，需要依托语料库完成，语料库设置需充分代表语言使用的广泛性。Biber（1990）的语料库研究表明，字（词）表构建所使用的语料库应包含不同题材、不同类型的文本，防止某一种文本的偏差产生不良影响。同时，使用语料的时间跨度也要足够长，以减少单独某一时期特色词和高频词的负面效果。

基于题材多样性和时间跨度的考虑，本研究选用文渊阁《四库全书》作为统计语料，构建了包含 3408 个古籍文本，25277 个字型，731852425 个字例的语料库用于字表研制。《四库全书》被誉为传统文化的巨典，古代典籍的渊薮（汪受宽、刘凤强，2005），分为经、史、子、集四部，在内容体裁上丰富多样，以其作为统计语料，可避免单独某一种类型文本对字表选字的负面影响，同时，它汇集了我国清代乾隆以前各朝代的主要文化典籍，将不同时期的用字情况纳入了考量，可以避免根据单独某个时期用字情况选取常用字的局限性。

2.2 选字数量

研制《古籍汉字常用字表》的目的，是为了让有古籍文献阅读需要的人花费最少的时间和精力优先学习一批用处最大、用得最多的字，而那些用处不大、用得很少的字，则较少考虑，可以不收入字表。

衡量一个字有用性的重要指标是字频（Nation，1997），即这个字在语言使用中出现的频率。齐夫定律表明，在英语单词中，只有极少数的词经常使用，绝大多数词使用的很少（MEJ Newman，2005）。对《四库全书》前 N 高频字型进行文本覆盖率测试，可以发现其用字同样符合这一规律。

前 N 高频字型数	覆盖率	前 N 高频字型数	覆盖率
前 500 字	67.8%	前 2500 字	94.9%
前 1000 字	82.1%	前 3000 字	96.3%
前 1500 字	88.9%	前 3500 字	97.3%
前 2000 字	92.6%	前 4000 字	98.0%

表 3. 前 N 高频字型所对应文本覆盖率

由表 3 可知，《四库全书》中前 1000 高频字型能实现对整个语料 82.1%的覆盖，而前 4000 高频字型也只能使覆盖率达到 98.0%。

由于古代识字教材“三百千”所收字型数均不超过 1000，参考其所收字型数及前 N 高频字型对文本覆盖率的情况，我们决定将字频排名前 1000 的字型选入《古籍汉字常用字表》进行考察，作为构建相关字表的初步尝试。

3 《古籍汉字常用字表》的分级

3.1 分级计量特征设计及标注方案

给古籍汉字分级是指确定各汉字的学习优先级别，解决哪些字先学、哪些字后学的问题。识字教学需考虑诸多因素，除了字频，还要结合汉字自身特点，合理设计学习顺序，做到“急需先学、先易后难、由浅入深、循序渐进”（李兆麟，2004）。因此，必须确定《古籍汉字常用字表》中各汉字的学习优先级别，研制分级字表，以期达到事半功倍的学习效果。

作为记录汉语的符号，汉字既有读音和意义，又有突出的字形形态。因此汉字分级计量特征的设计，需要围绕字音、字义、字形展开。同时，作为一种交际工具，在交际过程中，有的汉字使用得多些，有的使用得少些（冯志伟，1989），所以分级还需从汉字的应用层面考虑。现对各层面的分级计量特征及学习优先级别进行说明，在最终确定汉字等级时，各计量特征不是孤立的，需综合分析。

3.1.1 汉字的应用层面

字频：李国英、周晓文（2011）综合前人研究把汉字字频定义为“个体汉字字符在按特定原则选定的文本中出现的次数与选定文本总字次之比。”字频反映了一个汉字的常用度，在之前的工作中，我们把字频排名前 1000 的字型选入了《古籍汉字常用字表》。现把它们平均分为五组，各组的字频排名分别为 1-200、201-400、401-600、801-1000。分别赋予每组汉字 5、4、3、2、1 的学习优先级别，字频越靠前的，优先级别越高，学习优先级别由 5 到 1 递减。

使用度：确定一个字是否常用，不能单纯依靠字频，还须考虑其使用范围。如果某字出现的文本个数多，则说明其分布均匀，使用面广；反之则分布不均，使用面窄。我们用“使用度”表示汉字这一特征，其计算公式为：

$$\text{使用度} = \frac{\text{某字型出现的文本个数}}{\text{总文本个数}}$$

使用度越高，则该字型使用面越广，越需要优先学习。对 1000 字的使用度进行计算，结果如表 4 所示，绝大多数汉字的使用度都在 0.8 以上，占比达 92.2%，这些汉字分布均匀，使用面广，只有极个别汉字使用度在 0.6 以下，使用面较窄。

使用度	汉字个数	比例	累加比例
0.9-1	594	59.4%	59.4%
0.8-0.9	328	32.8%	92.2%
0.7-0.8	63	6.3%	98.5%
0.6-0.7	12	1.2%	99.7%
0.4-0.6	3	0.3%	100.0%

表 4.汉字使用度统计表

确定汉字学习顺序，需优先学习使用度高的汉字，因此需根据使用度对汉字进行分组，赋予各汉字不同的学习优先级别。如表 5 所示，由于 1000 字的使用度数值并非连续分布的，在[0.43,0.57]这一区间出现空缺，如果人工对其进行分组，各组别的取值范围将难以确定。因此，我们采用聚类分析的方法自动分组。聚类作为一种自动化程度较高的无监督机器学习方法，对某一具体的任务来说，分析之前数据所属的类别是未知的，聚类的目标就是将数据划分到不同类别中，同一个类中数据相似，不同类间数据相异。

如表 5 所示，以使用度数值为标准，采用 K-means 的聚类方法对 1000 字进行 5 组聚类，1000 字被划分到不同类别中。各组的取值范围分别是[0.36,0.43]、[0.57,0.77)、[0.77,0.86)、[0.86,0.93)、[0.93,1]，可见，聚类分析充分考虑到使用度数值在[0.43,0.57]这一区间缺失的特征，未将分组端点值选取在这一区间内。使用度越高的汉字越需要优先学习，据此赋予各组汉字 5、4、3、2、1 的学习优先级别。

聚类依据	类别	汉字个数	取值范围	学习优先级别
使用度	1	418	0.93-1	5
	2	349	0.86-0.93	4
	3	188	0.77-0.86	3
	4	43	0.57-0.77	2
	5	2	0.36-0.43	1

表 5.使用度聚类结果及学习优先级别

构词能力：汉字的构词能力是指汉字能否与其他汉字组合构成新词的能力，常用汉字参与构词的数量来衡量（江新，2006）。在考虑选择哪些汉字作为基础汉字时，除了要考虑字频外，还要考虑汉字的构词能力（赵金铭，1989）。从使用角度来看，一个汉字的构词能力强，则说明该字具有较强的实用性和组合能力，应当优先学习。

考察古籍汉字的构词能力，首先需要对古籍文本进行分词，构建词表。本研究选用程宁等（2020）的古汉语一体化词法分析软件对《四库全书》文本进行分词处理工作，该软件分词 F1 值达到 85.73%，效果较好。

由于《四库全书》文本数量众多，如果对所有的文本都进行分词，需要耗费大量时间，同时增加后期统计难度，因此，我们采用随机抽样的方法从 3408 个文本中选取了 200 个文本进行分词，获得包含 270034 个词型的词表，借助该词表考察 1000 字的构词能力。

尽管古汉语词汇具有以单音节词为主的特点（赵克勤，1987），但分词结果表明，《古籍汉字常用字表》选取的 1000 字，除了可以单独成词外，绝大多数还可以与其他字组合构成新词。掌握了构词能力强的字，便很容易认读理解它们组成的词语。以“王”为例，其参与构词有“魏王”、“国王”、“鬼戎王”等，学习者在习得“王”字的基本语义及用法后，在古籍阅读中，无论是遇到“陳王”、还是“楚王”，皆可推测其表示某君主或诸侯王。因此构词能力同样是赋予学习优先级别的一个指标。

3.1.2 字形层面

汉字是记录汉语的书写符号体系，是最重要的辅助性交际工具（李索，2004）。从书写符号的角度来看，汉字难度与其视觉呈现——字形密切相关。

字形结构：不同于线性排列的拼音文字，结构方式是汉字字形的重要特征，对汉字及其部件认知有一定的影响，彭瑞祥等（1983，1984）在母语者辨认不同结构汉字的研究中发现，左右（横向）结构字的再认率明显高于其他结构字，横向结构对于母语者来说最容易掌握，半包围结构的字较难再认；喻柏林等（1992）证明在心理切分上，上下结构字要极大地难于左右结构字。

对 1000 字的字形结构信息进行统计，发现主要分为两大类：由单个部件构成的独体字和由多个部件构成的合体字，合体字可以分为：左右结构（包括左中右结构）、上下结构（包括上中下结构）、包围结构（包括全包围结构、半包围结构），根据前人对不同结构汉字识别难度的研究，我们分别对 1000 字赋予结构学习优先级别，难度低的，学习优先级别较高,独体字、左右结构、上下结构和包围结构的汉字，学习优先级别分别为 4、3、2、1。

部首：不同的部首具有不同的构字能力。构字能力强的部首，如“氵”，具有较多的同部首字，如江、河、湖、海。因此，在掌握“氵”一个部首后，相当于对多个该部首字有所了解。

对 1000 字进行分析，共得到 192 个不同的部首，其中，29.7%的部首只能构成单个汉字，构字能力最弱，50.5%的部首构字数不超过 2 个，构字数量超过 10 个的部首较少，约占部首总数的 10%，在所有部首中，构字能力最强的是“口”，构字数多达 43 个。

对于那些包含强构字能力部首的汉字，应优先学习，例如部首为“言”的汉字“謂”、“諸”、“記”，在学习它们时，可以联想到其同部首字，加深印象，降低记忆负担（吴鑑城等，2019）。以构字数量为标准，对得到的192个部首进行聚类，分为五组，对包含对应部首的汉字赋予了不同的学习优先级别。

笔画数量：合体字由部件构成，独体字由笔画构成，但不论是合体字还是独体字，均可以分析到笔画。不同于拼音文字有长度上的差异，一个汉字无论简单还是复杂，都只占据一个方块的书写空间，其不同往往体现在笔画上。因此，笔画是汉字的最小结构单位。汉字笔画的多少，标志着该字视觉形状上的复杂程度（沈烈敏，1994）。不少研究表明，汉字的笔画数影响汉字识别，即存在笔画数效应，如曹传咏和沈晔（1963）。叶重新和刘英茂（1972）认为多笔画字的认识阈最高，最难认识，中笔画字次之，少笔画字认识阈最低。由此，我们推断：对于一个汉

字，其笔画数越多，则该字难度越大。

利用汉典网[1]，借助网络爬虫技术对 1000 字的笔画信息进行获取，分析得到笔画数量分布图：

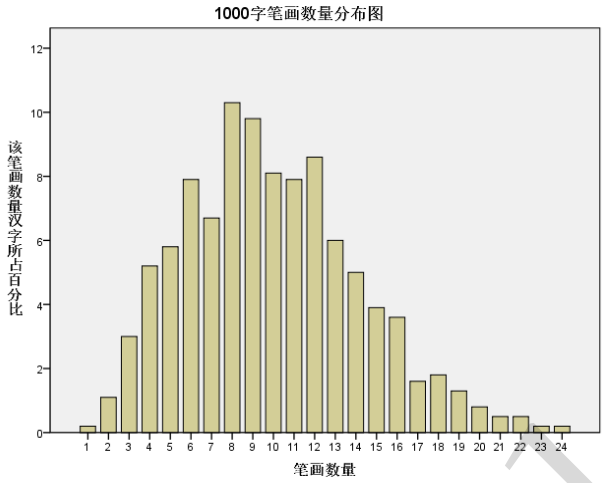

图 1.1000 字笔画数量分布图

从字形上看，笔画数越少的汉字越简单，越应优先进行学习，而笔画数多的汉字，书写较复杂，学习顺序应靠后。

3.1.3 字音层面

一个汉字可能有多个读音，相较于单音字，多音字发生误读进而产生理解错误的可能性更大，因此更难掌握。杨华（2003）发现，多音字读音数量上的差别不是造成其误读的主要原因，不同读音语用频率的差别才是影响其误读的主要因素。因此，从字音层面考虑多音字学习优先级时，不仅要考虑其读音数量，还要考虑其读音的语用频率。

对1000字的读音数量进行统计，我们发现绝大部分汉字为单音字，占比80.1%，在多音字中，双音字占比最高，为16.5%，三音字等总计占比3.4%。将单音字放在多音字之前学习是合理的，但多音字该如何处理呢？由于目前尚不能获得多音字中不同读音的语用频率，我们采取简化标注方案，不考虑多音字的读音数量对学习优先级别的影响，直接将汉字分为单音字和多音字，其中单音字的学习优先级别为2，多音字的学习优先级别为1，由此获得1000字在读音数量这一层面的学习优先级别。

3.1.4 字义层面

字的词性（用法）标签：古汉语词汇具有单音节词为主的特点，从书写符号的角度来看，一个个词语就是一个个汉字。在字典中，这些汉字不仅仅说明了字义，还标明了词性及用法。由于汉语缺乏形态变化，其词类和句法成分之间不存在简单的一一对应关系，一个汉字往往具有多个词性（用法）标签，这些标签是我们对汉字进行分级的重要依据。

借助汉典网上“國語辭典”对汉字的解释，我们对 1000 字的词性（用法）标签进行了获取，汉字的词性（用法）标签共分为名、动、形、代、副、叹、连、助、缀 10 类，一个汉字的（词性）用法标签个数越多，则说明它的用法越多，在实际使用中，该字能充当多种角色，应当优先学习。

义项个数：在汉字的每个词性（用法）标签下面，对应着多个不同的义项。如“書”字，在“动词”这一词性标签下，有两个义项：（1）写；（2）记载。义项个数能够说明字义的多样性。

[1] 汉典网（https://www.zdic.net）是一个免费在线辞典，收录了 93898 个汉字的信息。

一个汉字，其义项个数越多，能表示的含义也就越多，优先级别也就越高，应当优先学习。

以上，我们分别从字的应用、字形、字音、字义四个层面介绍了汉字分级的计量特征，现以各计量特征为标准，采用聚类分析的方法对 1000 字进行划分，获得各汉字在每一计量特征上的学习优先级别，结果如表 6 所示：

计量特征 / 取值范围 / 学习优先级别	构词数	字形结构	部首构字数	笔画数	读音数	词性用法标签数	义项数
1	2-417	包围结构	1-6	18-24	2-5	1	0-4
2	420-1020	上下结构	7-16	13-17	1	2-3	4-8
3	1037-1809	左右结构	24-25	9-12		4-5	9-12
4	1876-2834	独体字	31-34	6-8		6	13-18
5	3323-4539		40-43	1-5		7	19-24

表 6.不同计量特征下的汉字学习优先级别及特征取值范围

一个汉字可以用其计量特征及对应学习优先级别表示，例如： **難** = 字频 4+使用度 5+构词能力 1+笔画数量 1+字形结构 3+部首 2+字音 2+词性用法标签 3+义项 3

基于此，本文提出一种字向量模型，每个汉字由一个维度为 9 的向量表示，一个计量特征代表一个维度，对应维度的权重为该计量特征的学习优先级别，例如“**難**”字可以表示为字向量 **難**：（4,5,1,1,3,2,3,3），这样便可获得基于计量特征学习优先级别表示的字向量。

汉字	向量表示
長	(5,5,4,4,4,1,2,3,5)
為	(5,5,1,3,4,2,2,4,5)
齋	(1,3,1,2,1,1,1,2,2)
……	……

表 7.字向量举例

将各字向量映射到欧氏空间，进行基于学习优先级别的各汉字间相似度的计算。需要说明的是，我们的最终目的是为了实现汉字的分组，让相似度高的汉字聚集在一组，相似度低的汉字则聚集在另一组，而不是求得具体某汉字与其他汉字的相似值。因此，需要设置一个用于相似比较的标准。

假设存在这样一个**理想汉字**，它在字频、使用度、构词能力、笔画数量、字形结构、部首、字音、词性用法标签、义项等层面，均属于最优先学习一类，各层面的学习优先级别均为最高，则其可以表示为向量(5,5,5,5,4,5,2,5,5)，以该理想汉字为标准，那么与它相似度越高的汉字，越应该优先学习。

采用计算欧氏距离的方法测量各字向量与理想汉字间的距离，欧氏距离能够体现个体数值特征的绝对差异，适用于需要从各维度的数值大小中体现差异的分析，符合本研究的计算要求。在 m 维空间，点 x 与点 y 的欧氏距离的计算公式为:

$$\mathbf{D}(\boldsymbol{x},\boldsymbol{y})=\sqrt{\sum_{i=1}^{m}(x_m-y_m)^2}$$

求得各汉字与理想汉字的欧式距离用于后续汉字分级，结果如下：

汉字	与理想汉字的欧式距离
本	3.32
長	4.69
驚	10.10

表 8.汉字与理想汉字欧式距离举例

3.2 分级界标设置

分级意味着制造差别，同一级别内部成员应当是相似的，不同级别间的成员则存在着差异。如果直接主观地对汉字划分等级，差别可能不明显，必须依据某一特征量，才能使得分级有据可依。字表分为几级，要根据字表的需求来确定。以往的分级字表，一般分为 3-5 级，如《通用规范汉字表》(3 级)、《汉语国际教育用音节汉字词汇等级划分》(3 级)、《汉字频率表》(5 级)。《古籍汉字常用字表》目前仅收录 1000 字，数量较少，本身就是古籍阅读中应该掌握的基本汉字，因此划分级别数无需过多，我们确定为 3 级，更加突出重点，强调优先级别。以各汉字与理想汉字的欧式距离为依据，欧式距离越小，则与理想汉字越相似，越应优先学习，采用 K-means 聚类方法对 1000 字进行聚类，聚类数为 3，聚类结果如下：

类别	欧式距离取值范围	汉字个数
1	3.32-5.83	105
2	5.92-7.21	340
3	7.28-10.09	555

表 9.学习优先级别聚类结果

由此确定了共分三级的《古籍汉字分级字表》，其中一级字 105 个，二级字 340 个，三级字 555 个，一级字优先级别最高，最应优先学习，二级字、三级字优先级别递减。

4 《古籍汉字常用字表》与其他字表的对比分析

将《古籍汉字常用字表》与其他字表进行对比分析，可以帮助验证其收字是否合理。因此我们分别选用传统识字教材“三百千”和《现代汉语常用字表》与其进行比较。

识字课本	重合字型数	重合率
《三字经》	387	72.7%
《百家姓》	248	49.5%
《千字文》	604	60.4%

表 10.《古籍汉字常用字表》与“三百千”所收字型比较

表 10 展示了《古籍汉字常用字表》与“三百千”收字比较情况，可以看到《古籍汉字常用字表》与《百家姓》所收字型重合率不高，仅为 49.5%，这是由《百家姓》本身的内容所决定的。作为一本姓氏汇总读物，《百家姓》将姓氏作为选字依据，而不是从汉字本身特点出发进行选字汇编。张志公（1992）曾评价：《百家姓》里的字都是姓，儿童只要念这些字，认这些字的模样就行，无需去追究字义和句义。作为一本识字教材，《百家姓》的收字未考虑到汉字字形、字音等层面的具体特征，本身就具有较大的局限性，所以与《古籍汉字常用字表》字型重合率不高。

《三字经》和《千字文》与《古籍汉字常用字表》的字型重合率分别为 72.7%、60.4%，这

是因为《三字经》和《千字文》也不是完全地从识字教学角度进行选字，而在很大程度上考虑对儿童进行知识和思想教育的需要，因而更注重其内容组织的丰富性。《千字文》全书共 250 句，每 4 字一句，4 句一组，内容涉及天文地理、历史政治、封建纲常、伦理道德等各个方面，同时，为保证读起来朗朗上口，还要注意韵律，每两句一押韵。这种兼顾内容和用韵的文本内容组织，必然导致其在选字上不能完全从汉字本身特点出发。因此，与《古籍汉字常用字表》的字型重合率也不高。

《古籍汉字常用字表》 独有字	《千字文》 独有字	《古籍汉字常用字表》 与《千字文》共有字
一 十 三 至 山 又 今	羌 遐 盤 鬱 邇 駒 場	正 定 何 射 枝 習 日
未 太 六 矣 元 氏 風	賴 髮 鞠 毀 效 罔 談	王 重 仁 有 感 上 甚
北 前 七 里 小 江 凡	緣 璧 競 竭 履 溫 清	推 能 戶 寧 桓 尺 令
注 請 凡 郎 吏 殺 許	淵 澄 篤 慎 基 優 攝	弗 金 外 孟 遠 自 都
……	……	……

表 11.《古籍汉字常用字表》与《千字文》独、共有字情况

《古籍汉字常用字表》与《千字文》均收 1000 个字型，更适合对比分析。表 11 展示了《古籍汉字常用字表》与《千字文》的部分独有字与共有字。可知，单从字形这一层面考虑，《千字文》的独有字就不太简单，不适合儿童学习，如“髮、鞠、毀、緣、璧、競、攝”等，而像“一、十、三、至、山、又、然、今、未、太”这类字形简单的汉字，《千字文》却未收录。

对《千字文》和《古籍汉字常用字表》所收独有字进行考察：

	覆盖率	平均使用度	平均笔画数
《千字文》独有字	2.99%	0.35	12.6
《古籍汉字常用字表》独有字	19.10%	0.76	10.3

表 12.《古籍汉字常用字表》与《千字文》独有字统计信息

可见，《古籍汉字常用字表》所收独有字对《四库全书》的文本覆盖率和平均使用度更高、平均笔画数却更少，说明它们更常用、使用范围更广，在书写上更容易。因此，《千字文》中收录的部分汉字，合理性有待商榷。

《现代汉语常用字表》是现代汉字规范的重要字表，其所收汉字在很大程度上代表了现代汉字运用的基本情况。将《古籍汉字常用字表》与其进行比较，可以帮助我们比较分古籍文本与现代汉语常用汉字的异同。

对比分析发现，《古籍汉字常用字表》中，共有 652 字在《现代汉语常用字表》中出现，而未出现的 348 个字，均为繁体字，由于《现代汉语常用字表》中收录的都是经过简化的简体字，两者自然不能对应起来。

我们人工对这 348 字进行了繁简体转换，将其中的 316 个繁体字形转化为简体字形，再次与《现代汉语常用字表》进行比对，结果表明，经过简化的汉字中，有 311 字为两表共有字，因此，两个字表共有 964 字重合，这说明汉字系统具有极强的稳定性，96.4%的古籍汉字常用字至今仍为现代汉语常用字，它们很好地传承了下来，是汉字系统中的核心字。

而《古籍汉字常用字表》中独有的 36 个古籍汉字，具体包括：

《古籍汉字常用字表》独有字								
曰	郡	祀	虞	弗	襄	厥	吾	惟
羣	詔	丞	諭	諫	佐	庚	陛	兮
矣	哉	汝	朕	禹	耶	嗣	桓	録
焉	絶	闕	朔	仕	尧	巳	雍	蔡

表 13.《古籍汉字常用字表》独有字

分析可知，这些汉字未被《现代汉语常用字表》收录是有原因的，如古籍文本中表示说话的“曰”，表语气的“矣、哉、兮、耶、焉”，人称代词“吾、朕、汝”，这些字在现代汉语中，皆不再常用，而一些特殊名词：如姓氏“蔡”，常用来表人名的“禹、朔、尧、桓、襄”，表地名的“郡”，表官名的“仕、丞”、表天干地支的“巳、庚”，表君主尊称“陛下”的“陛”，颁布圣旨用的“詔、諭”以及臣子进谏的“諫”等，也因历史原因渐渐减少使用或逐渐废弃不用，这些古籍汉字未出现在《现代汉语常用字表》中是十分正常的。

5 结论与展望

不同于以往的字表研制，本论文在大规模古籍文本语料的基础上，考察了古籍文本用字信息，统计构建了《古籍汉字常用字表》，将其收字与传统识字课本“三百千”和《现代汉语常用字表》进行了比较，并在此基础上挖掘汉字分级计量特征，对字表中的汉字进行了宏观定量研究，考察了其字频、使用度、笔画、部首等信息。通过综合分析，对其中的汉字进行分级，进一步实现了《古籍汉字分级字表》的研制。然而，本研究仍有许多不足：首先，分级字量较少，基于目前的工作进度，我们只选择了古籍文本语料库中字频靠前的 1000 字进行了分级；其次，在利用汉字各层面计量特征时，未考虑到它们对汉字等级划分是否具有不同权重以及交互作用，而是无差别的平等对待；最后，《古籍汉字分级字表》的分级效果有待检验，需进行后续验证。在接下来的工作中，我们将针对以上问题，扩大分级字量、改进分级方法，进一步丰富完善《古籍汉字分级字表》的研制工作。

参考文献

Biber, D. 1990. *A typology of English texts.* Linguistics, 27: 3-43.

Johns J L. 1970. *The Dolch basic word list—Then and now*. Journal of Reading Behavior, 3(4): 35-40.

MEJ Newman 2005. *Power laws, Pareto distributions and Zipf's law* Contemporary Physics, 46:5, 323-351.

Nation P, Waring R. 1997.*Vocabulary size, text coverage and word lists*. Vocabulary: Description, acquisition and pedagogy, 14: 6-19.

陈黎明,张晗.“三百千”的用字及其流向[J].汉字文化,2010(01):57-62.

程宁,李斌,葛四嘉,郝星月,冯敏萱.基于 BiLSTM-CRF 的古汉语自动断句与词法分析一体化研究[J].中文信息学报,2020,34(04):1-9.

费锦昌.常用字的性质、特点及其选取标准[J].语文学习,1988(09):32-34.

“汉字应用水平测试研究”课题组,孙曼均.汉字应用水平测试用字的统计与分级[J].语言文字应用,2004(01):63-70.

江新,赵果,黄慧英,柳燕梅,王又民.外国学生汉语字词学习的影响因素——兼论《汉语水平大纲》字词的选择与分级[J].语言教学与研究,2006(02):14-22.

李国英,周晓文.汉字字频统计方法的改进[J].北京师范大学学报(社会科学版),2011,000(006):45-50.

李兆麟.谈常用字词的选取及其等级划分[J].辞书研究,2014(02):21-28.

彭瑞祥,张武田.速示下再认汉字的某些特征[J].心理学报,1984(01):49-54.

沈烈敏,朱晓平.汉字识别中笔画数与字频效应的研究[J].心理科学,1994(04):245-247.

汪受宽,刘凤强.《四库全书》研究的回顾与思考[J].史学史研究,2005(01):62-66.

吳鑑城,白明弘,林慶隆.臺灣華語文語料庫在華語文教育的應用[J].華語文教學研究 2019(03):29-56.

杨华.多音误读与语用频率的关系[J].语言文字应用,2003(02):30-38.

叶重新、刘英茂.影响本国文字认识阈的因素[R].台北:台湾大学心理学系研究报告,1972(14):113-117.

喻柏林,曹河析.汉字识别中的笔画数效应新探——兼论字频效应[J].心理学报.1992(02):120-126.

赵金铭.外国人基础汉语用字表草创[J].汉语研究,南开大学出版社.1989.

冯志伟.现代汉字和计算机[M],北京:北京大学出版社.1989.

国家汉语水平考试委员会办公室考试中心.汉语水平词汇与汉字等级大纲[M].北京:经济科学出版社.2001.

李索.汉字与中华传统文化[M].北京:高等教育出版社.2004.

彭瑞祥,喻柏林.不同结构的汉字再认的研究,普通心理学与实验心理学论文集[M].甘肃人民出版社.1983.

吴蒙.三字经 百家姓 千字文[M].上海:上海古籍出版社.1988.

苏培成.二十世纪的现代汉字研究[M],太原:书海出版社,2001.

孙钧锡.中国汉字学史[M].北京:学苑出版社,1991.

张志公.传统语文教育教材论[M].上海:上海教育出版社,1992.

赵克勤.古汉语词汇概要[M].浙江:浙江教育出版社,1987.