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.