基于深加工语料库的《唐诗三百首》难度分级

黄宇宇[1] ， 陈欣雨[1] ， 冯敏萱[12*] ， 王禹诺[1] ， 王蓓原[1] ， 李斌[12]
[1] 南京师范大学文学院
[2] 南京师范大学语言大数据与计算人文研究中心
hyuyuz@163.com

摘要

为辅助中小学教材及读本中唐诗的选取，本文基于对《唐诗三百首》分词、词性、典故标记的深加工语料库，据诗句可读性创新性地构建了分级标准，共分4层，共计8项可量化指标：字层（通假字）、词层（双字词）、句层（特殊句式、标题长度、诗句长度）、艺术层（典故、其他修辞、描写手法）。据以上8项指标对语料库中313首诗评分，建立基于量化特征的向量空间模型，以K-means聚类算法将诗歌聚类以对应小学、初中和高中3个学段的唐诗学习。

关键词： 《唐诗三百首》 ； 语料库 ； 难度分级 ； 诗句可读性 ； 文本长度

The difficulty classification of ' Three Hundred Tang Poems ' based on the deep processing corpus

Yuyu Huang[1] ， Xinyu Chen[1] ， Minxuan Feng[12*] ， Yunuo Wang[1] ， Beiyuan Wang[1] ， Bin Li[12]
[1] College of Arts Nanjing Normal University
[2] Center of Language Big Data and Computational Humanities
hyuyuz@163.com

Abstract

In order to assist the selection of Tang poetry in primary and secondary school textbooks and reading books, based on the deep processing corpus of word segmentation, part of speech and allusion markers of ' 300 Tang poems ', this paper innovatively constructs a grading standard according to the readability of verses, which is divided into 4 layers, a total of 8 quantifiable indicators : font layer (interchangeable words), word layer (double-word words), sentence layer (special sentence pattern, title length, verse length), art layer (allusions, other rhetoric, description techniques). According to the above eight indicators, 313 poems in the corpus are scored, and a vector space model based on quantitative features is established. The K-means clustering algorithm is used to cluster the poems to correspond to the Tang poetry learning of primary school, junior high school and senior high school.

Keywords: ' 300 Tang poems ' , Corpus , Difficulty classification , Readability of verse , Text length

*通讯作者
基金项目：江苏省社科基金项目(20JYB004)；古籍工作重点课题(22GJK006)；国家语委项目(YB145-41)；深圳爱阅基金会(儿童国学经典读物的分级阅读研究)

1 引言

唐诗数量丰富、艺术成就高，是中国优秀文化成果。据统计，我国的唐诗选本约有六百多种，而《唐诗三百首》是其中影响最大、流传最广、读者最多的选本，“风行海内，几至家置一编”，被视为唐诗入门启蒙读物首选(王东旭, 2013)。《唐诗三百首》是清乾隆蘅塘退士孙洙所编写的，本意是用来作为家塾读物训育儿童。《唐诗三百首》的选编理由并不只是诗歌的艺术成就，还考虑到了受众的接受情况、应试的需要、情感的“温柔敦厚”和教材的整体等，因此成就了《唐诗三百首》雅俗共赏、老少皆宜的高品位普及选本的地位。然而，《唐诗三百首》按体裁分章节进行编写，并未对诗歌进行难度分级，可能存在不符合学习者认知规律等问题。

根据学习认知规律，不同年龄段人群的阅读认知水平存在差异，学习积累古诗词是一个由繁到简的过程，不符合认知阶段的过难或过易的诗歌都可能会影响学习积极性，降低学习效果。而分级阅读是基于各个年龄阶段人群的认知水平，选择、提供适合不同年龄阶段阅读需要的文本。柳明烨(2021)因为阅读符合认知发展规律，因此有利于提高阅读古诗的效率，发展学习能力。

国内外对分级阅读的研究成果较为丰富，但是国内分级研究关注的多为现代汉语文本，对古代汉语文本的分级研究极少。针对文言文分级，张秋玲(2010)从2010年开始构建了文言文难易度评量的数学模型，首次为文言文难易度评量提出路径，经过两次修订，张秋玲等(2022)数学模型的相关系数提高了0.112，提高了数学模型评量文言文难易度的信效度。白瑞芬(2017)根据儿童心理学将分级阅读分为5个阅读阶段，提出了目前国学经典改编的问题和改进措施。针对古诗词分级，研究则更加少，柳明烨(2021)计算了课内外诗词的可读性，基于部编版小学语文教材的古诗词分级确立古诗词分级标准，对课外古诗词按照年级进行重新判别，具有开创性。但是对“诗句可读性”的指标制定颗粒度不够，不能较全面地覆盖特征。

因此，本文以赵昌平的《唐诗三百首全解》为底本(2006)，以基于条件随机场模型（CRF）的随园古汉语分词与词性标注系统深化加工处理，校勘与标注形成共计24540字、17896词规模的初始加工语料。后在此基础上标记唐诗的各类句式、修辞，建成《唐诗三百首》深加工语料库。基于该深加工语料库，本文量化影响可读性的维度，细化标注规则，设定特征权重，制定更为科学的古诗词分级标准和算法。

2 字层分级

2.1 通假字

《唐诗三百首》所使用的字都较为简单，生僻字较少。但是，唐诗诗句较为短小，距今时代久远，与目前的语言面貌不同，存在较多通假字、古今字和异体字现象，这些用例不符合当前人们的语言使用习惯，因此，字层面是影响《唐诗三百首》可读性的重要因素。

狭义的通假字与古今字和异体字有着千丝万缕的联系，但是并不等同。通假字是本有其字的假借，如“蚤”与“早”；古今字指同一个字在不同时代用不同字表示，如“莫”和“暮”；异体字指音义皆同、形体不同的两个字，如“泪”和“淚”(卢烈红, 2007; 李国英, 2007)。广义的通假字包括狭义通假字、古今字和异体字。本文为了便于标注和后续分级处理，使用的是广义通假字的定义。

影响通假字难度的因素包括出现次数和出现频率。出现次数指的是在一首古诗中，广义通假字出现的个数，出现个数越多，难度越大。出现频率指的是在古汉语文献中，该通假字的使用频率，使用频率越高，难度越小。赋分规则如下：出现次数上，一首古诗中出现N次通假字则计为N分。出现频率上，以使用频率为标准，将古汉语词义数据库中的通假字划分难度级别，难度分数为1-4分。再将出现次数和出现频率按照4：6的权重进行加权计算，根据所得分数将古诗分为0-4分。

3 词层分级

3.1 双字词

魏晋时期，汉字出现大量单音节词双音化的现象，发展到唐代，双字词数量较为丰富。双字词的语素按照组合规则划分，包括并列、偏正、主谓等关系，意义与两个语素也不尽然相同，双字词可能会发展出不同的意义。因此，理解双字词也是影响古诗阅读难度的重要因素，选定从语境的角度对诗歌这种较为封闭的文本中出现的双字词的特定语义理解进行难度分析。

第二十二届中国计算语言学大会论文集，第491页-第500页，哈尔滨，中国，2023年8月3日至5日。

立足于语境同语言的关系，可以分出“言内语境”(linguistic context)和“言外语境”(extra-linguistic context)两大类。言内语境，即文章或言谈中的话题的上下文或上下句，一般来说，对话语的理解依据是上文，听话人或读者对上文或上句作出推理，说话人然后又进一步说明，这种说明又成为听话人理解说话人意图的依据。各种语言语境的正确把握，对正确理解话语有着重要的作用。在有些场合，较小的语言环境不能解决问题，必须考虑较大的非语言环境。非语言环境指话语所发生的语言之外的环境，非语言环境也可称为情景上下文，它从各个方面影响着词的意义，如社会背景、语言情景、具体事件以及讲话方式等等。对各种言外语境的正确把握，对正确理解话语有着重要的作用。

在对《唐诗三百首》双字词进行理解难度的标分时主要以是否需由要“言外语境”和需要“言外语境”的程度进行0、1、2三个分数阶段的划分。其中分数0是指诗中出现的双字词在“言内语境”即可较为通顺的进行意义上的理解，不需要“言外语境”即可理解的双字词，这里所指的“言内语境”也就是诗中话题的上下文或上下句，一般来说，对话语的理解依据是上文，听话人或读者对上文或上句作出推理，说话人然后又进一步说明，这种说明又成为听话人理解说话人意图的依据。各种语言语境的正确把握，对正确理解话语有着重要的作用。分数0在三个分段中是数量最多的，共计2865个；分数1的双字词则是需要“言外语境”辅助理解，多有基于“言外语境”的引申含义只依靠“言内语境”会对造成理解偏差歧义，在有些场合，较小的语言环境不能解决问题，必须考虑较大的非语言环境。非语言环境指话语所发生的语言之外的环境，非语言环境也可称为情景上下文，它从各个方面影响着词的意义，如社会背景、语言情景、具体事件以及讲话方式等等，共361个；分数2的双字词是基于“言内语境”基本不能达意，必须借助“言外语境”进行理解的双字词，共56个。

Table 1: 《唐诗三百首》双字词各等级数量

双字词等级描述	等级划分	《唐诗三百首》中双字词数量
在“言内语境”即可较为通顺的进行意义上的理解	0	2865
需要“言外语境”辅助理解	1	361
基于“言内语境”基本不能达意，必须借助“言外语境”进行理解的双字词	2	56

4 句层分级

4.1 特殊句式

文言特殊句式，一般指的是文言文中不同于现代汉语表达习惯的某些特殊的句式，主要有判断句、被动句、省略句和倒装句等。唐诗诗句字数较短，每句多在5-7个字之间，且受到篇幅和格律的限制，出现省略句和倒装句的用例较多。另外，省略句省略句子成分，倒装句改变句子顺序，对唐诗可读性的影响较大。故将特殊句式中省略句和倒装句作为难度分级的指标。

常见的省略句类型有五种，包括省略关联词、省略介词、省略动词、省略比喻词和互文省略。前两种省略类型对意义影响不大，如“我歌月裴回，我舞影零乱”是“我歌（而）月裴回，我舞（而）影零乱”（李白《月下独酌》）的省略关联词形式。省略动词对意义影响一般，如“经冬犹绿林”是“经冬犹绿（满）林”（张九龄《感遇四首二》）的省略动词形式。后两种省略类型对意义影响很大，如“万事随转烛”是“万事（如）随转烛”（杜甫《佳人》）的省略比喻词形式。常见的倒装句类型有五种，包括主谓倒装、宾语前置、状语后置、定语后置和主宾倒装。所有倒装句类型对意义影响一般，如“碧玉妆成一树高”是“碧玉妆成一高树”（贺知章《咏柳》）的定语后置倒装。《唐诗三百首》特殊句式如图1所示。

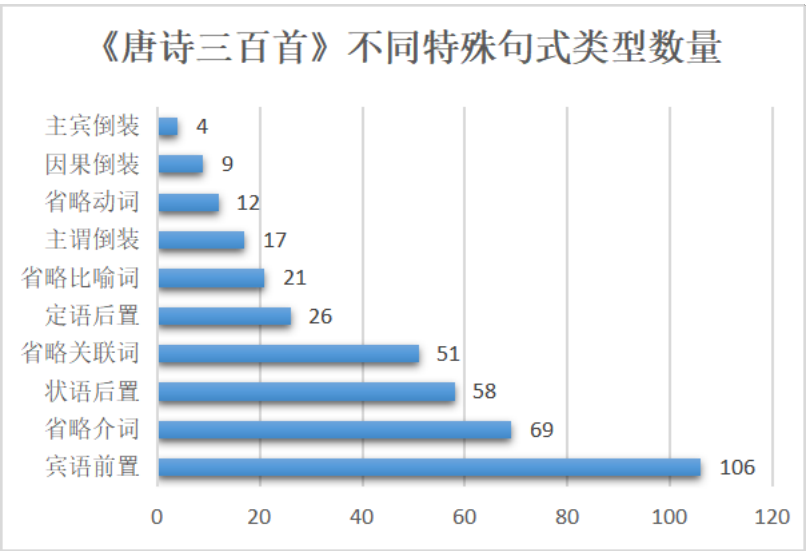

Figure 1: 《唐诗三百首》不同特殊句式类型的数量图

影响特殊句式难度的因素包括出现次数和对意义理解的影响程度。出现次数指的是在一首古诗中，特殊句式（省略句和倒装句）出现的个数，出现个数越多，难度越大。对意义理解的影响程度指的是该种句式类型（如省略关联词）对意义理解的影响程度，影响越大，难度越大。赋分规则如下：出现次数上，一首古诗中出现N次特殊句式（省略句或倒装句）则计为N分。对意义理解的影响程度上，将影响程度不大的类型计作1分（省略关联词和省略介词），将影响程序一般的类型计作2分（省略动词以及倒装句的五种类型），将影响程度很大的类型计作3分（省略比喻词和互文省略）。再将出现次数和对意义理解的影响程度，按照4：6的权重进行加权计算，根据所得分数将古诗分为0-4级。

4.2 诗句长度

为了便于统计，本文的诗歌长度指：除标题以外，一首诗歌的字数（不含标点符号）。《唐诗三百首》以体裁分章节，而体裁规定了每句诗的长度和句子数量，进而形成了每首诗的诗歌长度，诗歌长度对诗歌可读性影响很大。统计《唐诗三百首》的诗歌长度，如图2所示。可以发现，诗歌多为四种诗歌长度类型：20字（37首）、28字（60首）、40字（91首）、56字（55首）。从体裁来看，20字的诗歌为五言绝句（29首）和五绝乐府（8首），28字为七言绝句（51首）和七绝乐府（9首），40字为五言律诗（80首）、五言古诗（9首）和五古乐府（2首），56字为七言律诗（53首）、七言古诗（1首）和七律乐府（1首）。

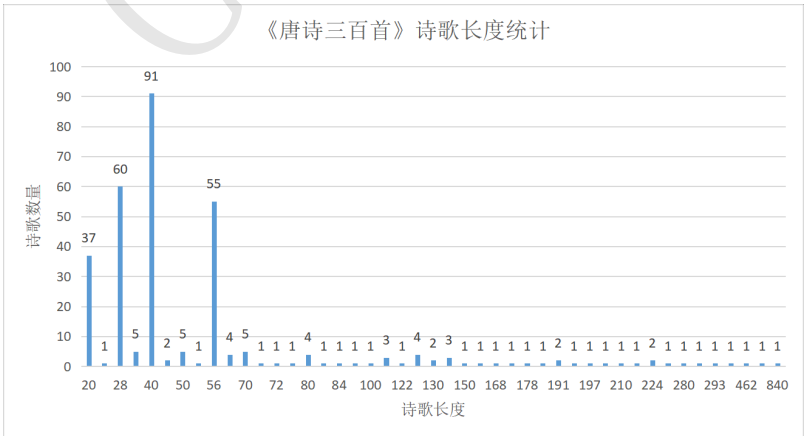

Figure 2: 《唐诗三百首》诗歌长度统计

再对人教版小学语文教材中唐朝诗人所写的古诗进行统计，如表2所示，小学低年级课文

诗歌长度多为20字，小学中年级所选的诗歌长度多为28字，小学高年级仍以20字和28字古诗为主，出现了部分40字和56字古诗。

Table 2: 人教版小学语文教材唐朝古诗诗歌长度统计

人教版小学语文教材唐朝古诗诗歌长度统计			
年级段	诗歌长度	诗歌数量	占该年级段比重
小学低年级（一二年级）	18	1	5%
	20	13	68%
	28	5	26%
小学中年级（三四年级）	20	4	24%
	28	13	76%
小学高年级（五六年级）	20	4	24%
	28	9	53%
	40	3	18%
	56	1	6%

基于对《唐诗三百首》中的诗歌长度分布统计和小学阶段唐诗诗歌长度的统计，本文将20字及以下的诗歌划分为1分，将21-28字诗歌划分为1.5分，将29-40字诗歌划分为2分，将41-56字划分为2.5分，将57字-100字诗歌划分为3分，将101-200字划分为3.5分，将200字以上划分为4分。

4.3 标题长度

标题长度对诗歌难度也有一定影响，标题长，理解难度越大。对《唐诗三百首》的标题长度（不含标点符号）进行统计，得到如图3结果。诗歌标题长度在1-50个字之间，并集中在2-6个字之间，标题长度在2-6个字诗歌占《唐诗三百首》总数量的比重为76%。

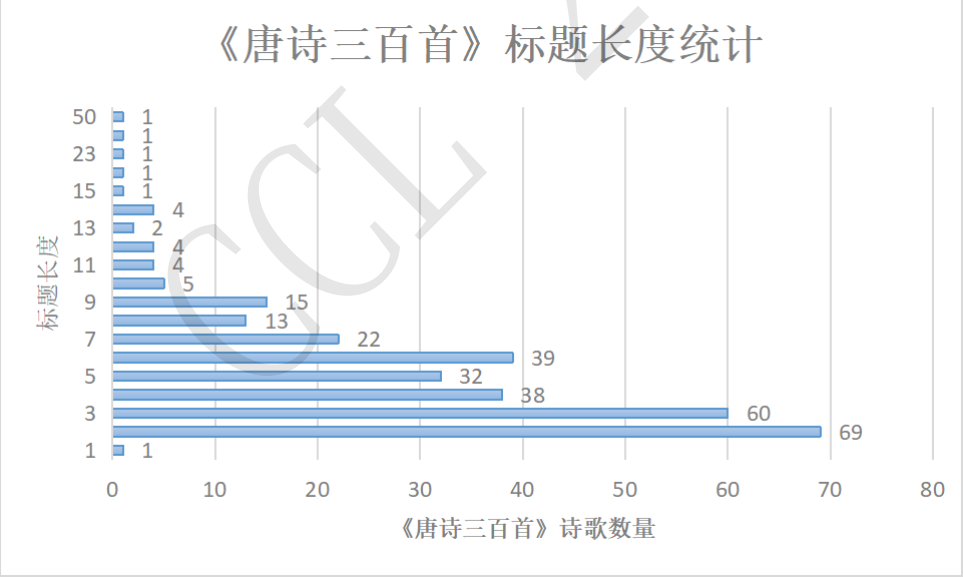

Figure 3: 《唐诗三百首》标题长度统计

再对人教版小学语文教材中唐朝古诗的标题长度进行统计。小学低年级的标题长度在1-7个字之间，标题长度为2个字的最多。小学中年级的标题长度在1-9个字之间，标题长度为2个字的最多。小学高年级的标题长度在1-10个字之间，标题长度为4个字的最多。基于《唐诗三百首》的标题长度统计和人教版小学语文唐朝古诗的诗歌长度，设计了标题长度的分级标准。标题长度为1-4个字，为1分；标题长度为5-6个字，为2分，标题长度为7-10个字，为3分；标题长度为10个字以上，为4分。

5 艺术层分级

5.1 典故

“典故”在《现代汉语词典》（第七版）的解释是“诗文等所引用的古书中的故事或词句。”也就是说，典故是古诗文中作者引用的古代故事或有出处的词句。从典故的内容角度进行分类，典故可以分为“事典”和“语典”。“事典”指的是在诗文作品中引用的事件性典故，通常包括上古神话、传说故事、历史故事、宗教故事等。“语典”指的是在诗文作品中引用的语言性典故，如前人所说的能够溯源的话语、诗词曲赋中的词汇短语、文学作品中的成语俗语等。从引用典故的方式角度进行分类，可将典故分为明用和暗用两类。明用典故是指作者在进行文学创作时，直接引用历史典故或简单概述历史故事。暗用典故是指作者在行文中对于典故的使用比较隐蔽，不像明用典故那样有直接的引用痕迹，而是通常将引用的典故消融在作品的字里行间，曲折内敛地表述情感。

在标注典故的过程中，以赵昌平《唐诗三百首全解》(2006)作为主要参考，并辅之以唐诗三百首语料库和搜索引擎，遵循尽可能多标注的原则，将典故全部标出。典故标注内容包括六类，分别为典故数量、典故类型、出处朝代、出处典籍和典故内容。标注结果显示：《唐诗三百首》收录的314首诗中共有71首诗用典，其中每首诗的用典数量如图4所示，其中一首诗中最多的典故数量为9个。

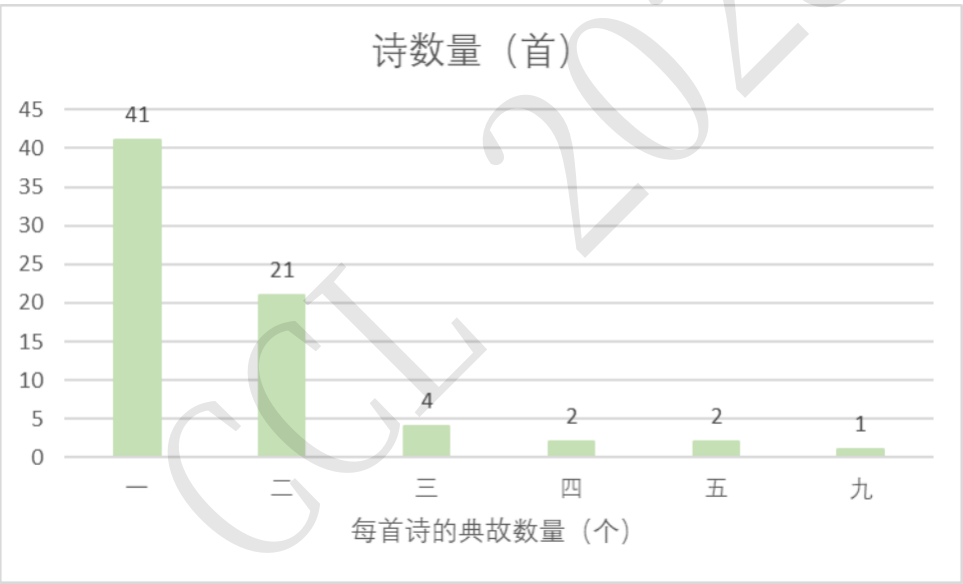

Figure 4: 《唐诗三百首》典故数量统计

影响典故难度的因素包括内容难度、出处难度和数量难度。第一，典故内容难度分类包括典故类型（事典和语典）和用典类型（明用和暗用）。根据试标结果，语典的流传度高于事典，理解难度上，事典＞语典；明用典故与暗用典故相比，更为直白，理解难度上，明用＞暗用。四个维度相较，暗用难度最高，因此对暗用赋以更高的分数。第二，出处难度包括出处朝代和出处典籍。根据BCC语料库统计出每个朝代和典籍出现的次数，出现次数越多，越容易理解，再根据次数分别划为四个等级。第三，数量难度指的是每首古诗出现的典故总数，典故越多，难度越大。三个维度的评分标准如表3所示。再将三个维度通过向量空间模型转化成特征值，设置在各个特征可读性难度均最高的理想诗句向量，计算诗句向量和理想诗句向量的欧式距离，从而得出典故的难度级别。

第二十二届中国计算语言学大会论文集，第491页-第500页，哈尔滨，中国，2023年8月3日至5日。

Table 3: 典故难度三维度评分标准

内容难度	**分数**	**语典**	**事典**
	明用	1	2
	暗用	3	4
出处难度	**典籍出现次数**	**分数**	
	1-100	1	
	100-500	2	
	500-1000	3	
	＞1000	4	
数量难度	**每首诗典故数量**	**分数**	
	1	1	
	2-3	2	
	4-5	3	
	9	4	

5.2 其他修辞

修辞手法是为提高表达效果，用于各种文章或应用文，在语言写作时表达方法的集合。古诗的修辞手法较为丰富，常见的如比喻、拟人、隐语和借代等。因为典故单独作为指标，此处其他修辞指除用典以外的修辞。与直接描写不同，修辞手法或增加喻体（如“大漠沙如雪，燕山月似钩。”——李贺《马诗》），如或将物拟人化（如“举杯邀明月，对影成三人。”——李白《月下独酌》），或用“借体”代替“本体”（如“朱门酒肉臭，路有冻死骨。”——杜甫《自京赴奉先县咏怀五百字》，不一而足。

影响修辞难度的因素主要是修辞数量，即一首诗歌中，修辞数量越多，阅读难度越大。另外，因为诗歌中使用比喻的修辞手法非常常见，数量很多，且不同喻体对诗歌可理解性影响很大，因此又根据对比喻中的喻体熟悉度，将其细分为三个难度级别（1-3级，从最不熟悉到最熟悉）。1级为最不熟悉，是在具体的情境中才可以理解的诗句，往往是暗喻或借喻（如“锦瑟无端五十弦，一弦一柱思华年”——李商隐《无题》）。3级为最熟悉，是在如今的日常语言中依然使用，或者很好理解的比喻（如“天边树若荠，江畔舟如月。”——孟浩然《秋登兰山寄张五》）。2级难度介于1级和3级之间（如“波澜誓不起，妾心井中水。天阶夜色凉如水，坐看牵牛织女星”——孟郊《列女操》）。

5.3 描写手法

描写就是作者对人物、事件和环境所作的具体描绘和刻画。描写方法：是用生动形象的语言把人物、事件、景物具体描绘出来的一种手法，给读者以身临其境的感觉。描写手法包括白描、象征、衬托、烘托、渲染等，另外诗歌中比较独特的描写手法包括比兴、主客位移等。

影响描写手法难度的因素包括描写手法数量和各个描写手法的理解难易度。描写手法数量越多，理解难度越大。各个描写手法的理解难度越高，诗歌整体难度越大。描写手法可以分为正面描写和侧面描写，侧面描写更委婉，理解难度高于正面描写。因此对属于正面描写的手法加权赋分为1分，侧面描写和特殊的描写手法（如主客位移）加权赋分为2分。

6 分级算法及验证

6.1 分级算法

本文利用AHP层次分析法计算上述8项指标的权重。AHP层次分析法（Analytic Hierarchy Process）是对于定性的决策问题进行定量化分析的一种方法。如表4，针对通假字、特殊句式、修辞、描写手法、典故、双字词、诗句长度、标题长度构建8阶判断矩阵进行AHP层次法研究，分析得到特征向量和权重值。结合特征向量可计算出最大特征根(8.299)，接着利用最大特征根值计算得到CI值(0.043)，结合判断矩阵阶数得到RI值，计算CR值（CR=CI/RI），并且进行一致性判断。本次针对8阶判断矩阵计算得到CI值为0.043，针对RI值查表为1.410，因此计算得到CR值为0.030＜0.1，意味着本次研究判断矩阵满足一致性检验，计算所得权重具有一致性。

Table 4: AHP层次分析法计算权重结果

AHP层次分析结果				
项	特征向量	权重值	最大特征值	CI值
通假字	0.35	0.0438	8.299	0.043
特殊句式	1.308	0.1635		
修辞	1.513	0.1891		
描写手法	1.513	0.1891		
典故	1.308	0.1635		
双字词	0.35	0.0438		
诗句长度	1.308	0.1635		
标题长度	0.35	0.0438		

利用AHP层次分析法设定好权重后，使用向量空间模型(VSM: Vector Space Model)，将诗句可读性的各个特征转换为标注体系，建立基于计量特征的向量空间模型。每首诗由一个维度为8的向量表示，一个计量特征代表一个维度，每一特征项都对应一个权重，对应维度的权重为该计量特征的学习优先级别。这样一个向量可用它含有的特征项及其特征项所对应的权重所表示。

并利用相似度计算的方法，设置在各个计量特征层面上，诗句可读性难度均为最高的理想诗歌（即八个维度都为最高分4分），则其各个维度均达到满分，以该理想诗歌为标准，计算其他诗歌与它的相似度，相似度越高的诗歌，理解难度越大，诗歌可读性越低。采用计算欧氏距离的方法测量各诗歌向量与理想诗歌间的距离。

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

计算完成后，对分值进行降序排序，得分越低的诗说明和难度最高的理想诗歌越接近，即难度越大，排在越后边，得分越高的诗说明和难度最高的理想诗歌距离越远，即难度越低，排在越前边。

分级意味着同一级别内部差异不大，与其他级别则存在较大的差异。《唐诗三百首》作为蒙学教材，可以按照学习阶段来区分，因此可以分为小学阶段、初中阶段以及高中阶段三个等级。以每首诗与理想诗歌的欧式距离为依据，采用K-means聚类算法对《唐诗三百首》进行聚类，聚类数为3，结果如表5所示。从表可以看出：最终聚类得到3类群体，其中小学阶段131首，初中阶段127首，高中阶段55首，此3类群体的占比分别是41.85%，40.58%，17.57%。整体来看，3个阶段的诗歌分布较为均匀，整体说明聚类效果较好。

Table 5: 不同类别诗歌及欧式距离取值范围

类别	欧式距离取值范围	诗歌数量
1	10.66±0.42	131
2	9.41±0.38	127
3	7.79±0.80	55

6.2 难度等级验证

根据分级算法，《唐诗三百首》难度最高和最低的前十首如图5、图6所示。

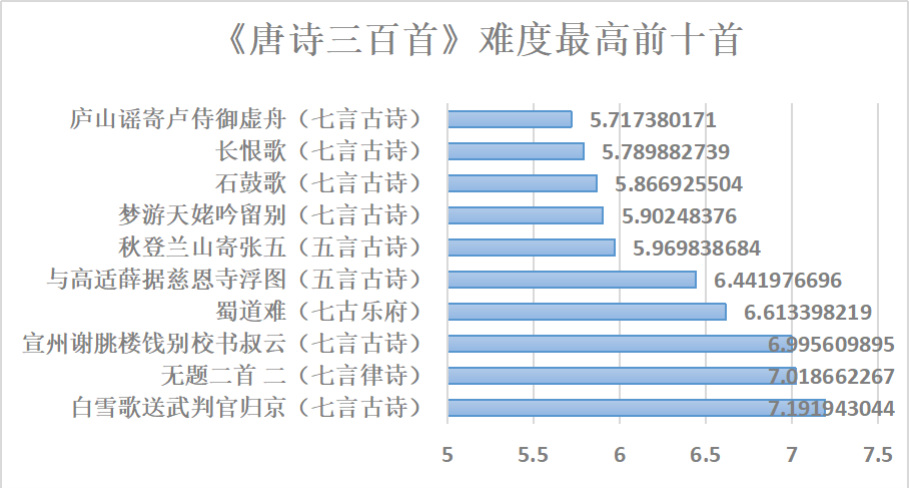

Figure 5: 《唐诗三百首》难度最高前十首

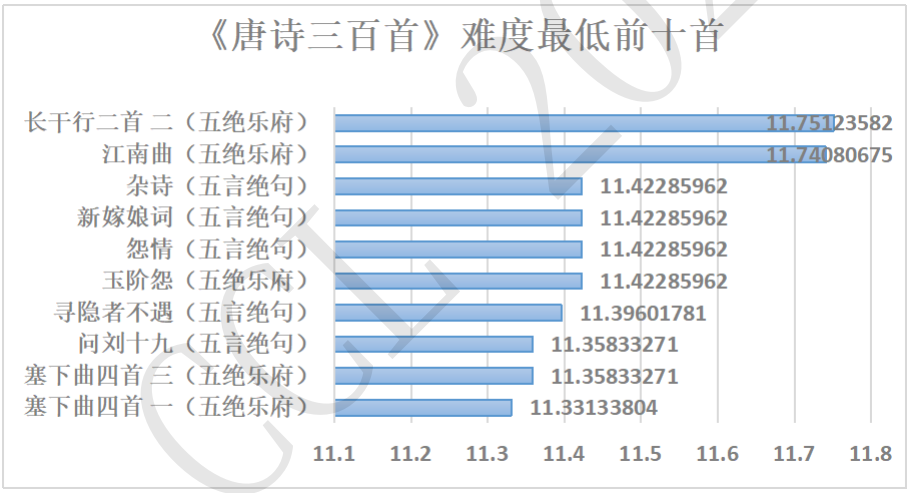

Figure 6: 《唐诗三百首》难度最低前十首

再将中小学阶段语文教材中出现的《唐诗三百首》诗篇与划分的等级进行比较，语文教材里小学阶段的诗歌，除了《山居秋暝》和《过故人庄》被划分在2类别（初中阶段），其他均在1类别（小学阶段）。教材初中阶段的诗歌，除了《宣州谢楼饯别校书叔云》、《白雪歌送武判官归京》和《兵车行》被划分为3类别（高中阶段），《送杜少府之任蜀州》被划分为1类别（小学阶段），其他均在2类别（初中阶段）。教材高中阶段的诗歌，如表6，有1篇被归为1类（小学阶段），有7篇被归为2类（初中阶段），有5篇被归为（高中阶段）。整体来看，分级结果与教材是贴合的。

第二十二届中国计算语言学大会论文集，第491页-第500页，哈尔滨，中国，2023年8月3日至5日。

Table 6: 不同类别诗歌及欧式距离取值范围

高中语文教材与《唐诗三百首》重合的诗歌篇目	诗人	体裁	分级分数	分级类别
《客至》	杜甫	七言律诗	10.34201827	1
《登高》	杜甫	七言律诗	9.96939296	2
《旅夜书怀》	杜甫	五言律诗	9.729834535	2
《燕歌行并序》	高适	七古乐府	9.708459972	2
《登岳阳楼》	杜甫	五言律诗	9.292056666	2
《将进酒》	李白	七古乐府	9.153705262	2
《琵琶行》	白居易	七言古诗	8.90408068	2
《蜀相》	杜甫	七言律诗	8.722441275	2
《无题》	李商隐	七言律诗	8.673868804	3
《锦瑟》	李商隐	七言律诗	7.944846913	3
《蜀道难》	李白	七古乐府	6.613398219	3
《梦游天姥吟留别》	李白	七言古诗	5.90248376	3
《长恨歌》	白居易	七言古诗	5.789882739	3

7 结论和展望

为了实现《唐诗三百首》的难度分级，本文在考察影响古诗阅读的因素之后，从4个层面设计了8个计量指标，并有针对性地细化标注规则，以诗歌为单位对《唐诗三百首》进行标注，形成向量空间模型，在利用AHP层次分类进行权重设定，最终利用K-means聚类形成3类难度级别。3类难度级别分别对应小学阶段、初中阶段和高中阶段，其中小学阶段有131首诗歌，初中阶段有127首诗歌，高中阶段有55首诗歌。经过与中小学语文教材进行验证，发现本论文分级算法是精密可行的。

然而，本研究仍有值得改进和发展之处，首先，各个指标权重设定时，主观判断的比重较大，因此仍需寻找各个指标重要程度的理论依据，或通过调查实践，了解各指标的重要程度。其次，虽然我们设计了8个指标，但是仍然有发展空间，可以扩充其他指标，如情感的分级。最后，分级阅读最终目的是推广阅读，在未来的工作中，我们会致力于将分级指标继续完善，助力古诗阅读的推广。

参考文献

卢烈红. 2007. 古今字与同源字、假借字、通假字、异体字的关系. 语文知识, (1):45–48.

张秋玲, 牛青森, and 赵宁宁. 2022. 中学语文教科书文言选文难易度评量模型检验. 语言文字应用, (3):49–61.

张秋玲. 2010. 文言文“浅易”的语词特征研究——以百年来初中教科书中的文言选篇为研究对象. 语言文字应用, (3):115–122.

李国英. 2007. 异体字的定义与类型. 北京师范大学学报*(*社会科学版*)*, No.201(46-50).

柳明烨. 2021. 基于部编版小学语文教材的古诗词分级标准研究. Ph.D. thesis, 南京师范大学.

王东旭. 2013. 论《唐诗三百首》与语文教材的古诗选编. Ph.D. thesis, 东北师范大学.

白瑞芬. 2017. 国家经典少儿读物分极改编的问题与思路. 编辑学刊, (3):102–106.

赵昌平. 2006. 唐诗三百首全解. 上海：复旦大学出版社.