Advances in Education 教育进展, 2023, 13(4), 1981-1990
Published Online April 2023 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2023.134312

Hans汉斯

国际中文教育教材与真题卷可读性对比研究

谭正娇

云南大学汉语国际教育学院，云南 昆明

收稿日期：2023年3月19日；录用日期：2023年4月16日；发布日期：2023年4月23日

摘　要

阅读是新HSK考试中非常重要的部分，也是学习者获取汉语知识的重要途径。通过对比课文与真题卷阅读部分的可读性，发现课文与真题卷的整体难度高度契合，但教材中存在课文间难度差异过大、难度增加缺少规律等问题，相较之下，真题卷表现出持续的稳定性。因此，教师在教学过程中应该注意从不同角度培养学生的阅读能力；学生应该主动使用不同的阅读技巧、从不同渠道获取阅读材料；教材编写者应避免课文间难度差异过大、积极寻求与其他网站的合作，减少课文语料存在滞后性这一问题。

关键词

教材，真题卷，可读性，对比

A Comparative Study on Readability of International Chinese Education Textbooks and Real Question Books

Zhengjiao Tan

School of International Chinese Language Education, Yunnan University, Kunming Yunnan

Received: Mar. 19th, 2023; accepted: Apr. 16th, 2023; published: Apr. 23rd, 2023

Abstract

Reading is a very important part of the new HSK test, and it is also an important way for learners to acquire Chinese knowledge. By comparing the readability of the reading part of the text and the real question paper, it is found that the overall difficulty of the text and the real question paper is highly consistent. However, there are some problems in the textbook, such as the large difference in difficulty between the texts, the increase in difficulty and the lack of rules. By contrast, the real question paper shows continuous stability. Therefore, teachers should pay attention to cultivating

文章引用: 谭正娇. 国际中文教育教材与真题卷可读性对比研究[J]. 教育进展, 2023, 13(4): 1981-1990.
DOI: 10.12677/ae.2023.134312

students' reading ability from different angles in the teaching process; Students should take the initiative to use different reading skills and obtain reading materials from different sources. Textbook writers should avoid the difficulty difference between texts, and actively seek cooperation with other websites to reduce the lag of text corpus.

Keywords

Textbook, Real Question Paper, Readability, Comparing

1. 引言

阅读是获取知识的重要手段和最佳途径之一。阅读理解题是一种综合性的题型，它能有效地检测学生的阅读理解能力和中文素养，阅读理解题历来是整个考试的重点与难点。在教材中，课文质量是衡量教材质量的重中之重。如何利用好一本教材，贯彻好“考教结合”理念，是每一个从事国际中文教育的人都应该认真思考的问题。

据中外语言交流合作中心统计[1]，2021 年上半年，全球约有 18 万余名考生参加了 HSK、HSKK、BCT 和 YCT 各级别考试，比去年同期增长近 50%，在俄罗斯，参加 HSK 五级、六级的考生人数显著增加。随着参加考试的人数增加，HSK 的重要性也比以往更加明显。HSK 成绩是外国学生申请到中国留学、申请各项中国奖学金的一项必要凭证；也是外国人申请在华工作、移民等的重要语言依据[2]。阅读是 HSK 的重要组成部分，分值占比约为 1/3。郑英等[3]开展了一项针对英语母语者的 HSK 考后感受调查，其调查结果显示：考生普遍认为在 HSK 中，到了中高级阶段(HSK3 级及以上)，阅读部分较其他部分更难得分。罗民等[4]调查发现，在 HSK 中阅读部分的难度高于听力部分，这些调查结果提示我们要加强对学习者 HSK 五级阅读能力与技巧的关注。

新 HSK 考试作为一项标准化考试，能检测出学生的真实汉语水平。《HSK 标准教程》作为北京语言大学专家团队编写出来的综合型教材，上市之后便受到了广大使用者的欢迎；《HSK 标准教程(5)》是整套教材的高级阶段课本之一，其编写形式除了沿用之前主题式的课文编排，还将课文的编写形式由之前的对话式改为短文式，这样更有利于学生通过教材了解 HSK 阅读部分的考察形式，让学生提前适应 HSK 的考试难度。因此，本文将以《HSK 标准教程(5)》(以下简称“《教程(5)》”)的课文与新 HSK 五级真题卷(以下简称“真题卷”)阅读部分为研究对象，通过自建语料库，将教材与真题卷进行可读性对比研究，以帮助教师与学生更好地利用教材。

2. 可读性与可读性公式

吕中舌([5], p. 117)提出可读性(readability)又被称为易读性，指一个文本对于读者来说是否易于阅读。通过对文本进行可读性分析并将阅读材料根据可读性分数划分等级，能够帮助读者选择难度适宜的阅读材料。特别是对于第二语言学习者来说，选择适合自己的阅读材料有利于第二语言的学习与发展，避免学习者选择过于简单的材料而无法提升自己的外语水平，也避免因选择的材料过难以至于消耗自己的学习热情。

可读性公式([5], p. 117) (readability formula)是指通过对文本中可能影响阅读体验的因素进行统计分析，最终制定出来可供测量文本难易程度的一种计量公式；可读性公式通过计量学的方式，将文本进行难易程度的分级。在可读性公式正式提出之前，即使人们已经意识到对文本进行可读性分析以及分级的重要性，但是由于没有统一的分析与统计标准，对文本分级大部分是依靠教师的教学经验，教师的经验与主观意愿在文本分级中发挥着很大作用，也由此导致对文本的分级不清晰、定位不清楚等问题。

王蕾[6]将可读性公式的发展分为三个阶段——上世纪 20 年代至 70 年代末为可读性公式研究范式形成阶段、70 年代末期至 90 年代末期为已成型可读性公式的实践验证阶段、90 年代末至今为可读性公式与其他学科融合发展阶段。周东杰、郑泽芝[7]认为，在国内的教育领域中，可读性研究可以分为母语教材、英语教材与阅读文本、对外汉语教材与读物、文学作品、翻译、科技论文几个方面，其中英语教材与阅读文本、对外汉语教材与读物是目前可读性的研究重点。

国际中文教育领域中，不管是可读性公式的开发还是运用都仍处于起步阶段相对来说，公式的实证性运用比开发更少。因此，在以后的研究中，一方面要加强对各类专门型公式以及可读性公式与机器学习相结合的研究；另一方面，公式研究出来之后要加强对其的运用，从实际运用中发现可读性公式存在的不足，以便于后续改进。本文将结合语料库语言学的研究方法与可读性公式的运用，一方面对比两类文本的可读性，另一方面也对已开发的可读性公式进行实证性运用。

3. 课本与真题卷的难度计算

为保证语料的真实性，本研究真题部分的语料采用的是已由语合中心(国家汉办)公开出版的 2012 版[8]、2014 版[9]、2018 版[10]《汉语水平考试真题集》。

本研究采用建立语料库的方式对语料的可读性分数进行计算，总的来说，可以将语料的可读性分数计算步骤分为以下几步：

3.1. 建库

将《教程(5)》中的课文与真题卷中的阅读部分分别建立语料库，建库过程中，将识别后的文本进行删除与修改。其中教材部分需要删除标题、图片、语言点、课后习题等内容，仅留下教材中的课文部分；真题卷部分删除听力、写作、阅读部分的插图，仅保留试卷中的阅读文本部分。将删除后的文档进行归类编码，如：教材上册第一课的课文编号为 0101，下册第一课的课文为 0201；针对 HSK 真题卷中已有编号的试卷，对原有编号不做更改；针对 2018 版真题卷中没有数字编号这一情况，遵循其他年份真题卷的编号方式，对其进行编号，编号为：H51801-H51805。

3.2. 分词

建立好语料库之后，利用哈工大 LTP 分词系统[11]进行分词处理。分词过程中，由于汉语分词标准不统一，我们采用“软件分词 + 人工校对”的方式对分词结果进行检查。检查过程中，重点检查超纲词、词汇划分错误等部分，并将修正后的结果重新放入语料库。

3.3. 数据统计

采用“汉语助研”软件统计语料中的字词数量进行统计[12]，然后根据郭望皓的可读性公式对两个语料库中的用字、用词、平均句长与可读性分数进行计算。其中用字部分需要统计语料中的字种，即一篇语料中不重复出现的字([13], p. 49)；用词部分需要统计语料中的某一级词的词次占总词次数的比例([13], p. 49)；平均句长指在语料中出现的所有语料被平均分为多少个句子，该部分以语料中的标点符号为判断依据。将基础数据进行统计处理后，再利用选定的可读性公式对语料的可读性进行计算。

3.4. 可读性分数计算

郭望皓[14]通过对以往可读性公式的考察，从汉字、词汇、句子三个维度对对外汉语文本的可读性进行分析，这三个维度能较大限度地顾及到文本构成单位的各个方面。具体来说，汉字层面从汉字笔画数与字频进行测定；词汇层面从词频、实词与虚词的占比、词长三个方面进行测定；句子层面从平均句长和篇长两个方面进行测定。作者通过计算机将公式进行拟合后对公式进行了验证使用，验证结果表明该公式的拟合度较高，后人对该公式的研究中也证明其适用于对长文本的可读性分数计算[15]。

以《教程(5)》上册第一课的课文语料为例，将语料进行上述处理后的结果见表 1：

Table 1. Example for calculating the difficulty of corpus words
表 1. 语料字词难度计算示例

语料来源	比较项目	甲级字字种	甲级字字次	乙级字字种	乙级字字次	丙级字字种	丙级字字次	丁级字字种	丁级字字次	纲外字字种	纲外字字次
0101	数量	196	489	48	91	5	7	4	6	0	0
	比例	77.47	82.46	18.97	15.35	1.98	1.18	1.58	1.01	0.00	0.00
语料来源	比较项目	甲级词词种	甲级词词次	乙级词词种	乙级词词次	丙级词词种	丙级词词次	丁级词词种	丁级词词次	超纲词词种	超纲词词次
0101	数量	129	258	47	65	21	43	8	12	29	42
	比例	55.13	61.43	20.09	15.48	8.97	10.24	3.42	2.86	12.39	10.00

根据统计出的字种数及占比，利用 0101 号语料中字种数量计算出文本的用字难度为：

$Y_{zi} = 0.148*196 + 0.182*48 + 0.137*5 + 0.215*4 + 0.283*0 = 39.289$

将语料的词次比代入用词难度的计算公式后，可以得出 0101 号语料的用词难度为：

$Y_{ci} = 0.132*0.6143 + 0.185*0.1548 + 0.249*0.1024 + 0.246*0.0286 + 0.188*0.1 = 0.1611$

0101 号语料的总字符数为 1640 个，语料中用于断句的标点符号总计为 26 个。则 0101 号语料的平均句长为：1640/26 = 63.08。在计算出语料的用字难度、用词难度与平均句长后，便可以将相关基础数值代入可读性公式并进行计算，因此 0101 号语料的可读性分数为：

$Y = -11.946 + 0.123*63.08 + 0.198*109.325 + 0.811*0.1611 = 17.5898$

通过将语料库中的语料代入可读性公式进行计算后，我们将各项计算结果绘制成散点图，结果见图 1 与图 2。

从散点图可以看出，教材与真题卷的用字难度呈现出如下特点：《教程(5)》的用字难度随学习时间的增加呈上升趋势，课文之间的用字难度、单篇课文与平均用字难度都存在较大差异。下册与上册相比，课文之间用字难度差异更大。在下册中，不仅是单元与单元之间存在这个问题，单元内的课文之间也有相同的问题。真题卷中的用字难度整体处于一个较为稳定的状态，虽然与用字难度平均值存在一定差异，但差距较小。其用字难度随时间推移而增加，呈缓慢上升趋势，这一现象在 2012 版与 2104 版真题集中表现得较为明显。

不管是课文语料还是真题卷语料，其用字难度与语料中的丁级字以及超纲字都呈正比。如果语料中的丁级字与超纲字占比高，则该语料的用字难度就高；相反，如果语料中丁级字与超纲字的占比低，则该语料的用字难度低。

其次，计算出教材与真题卷的用词难度，并绘制成散点图图 3 与图 4。

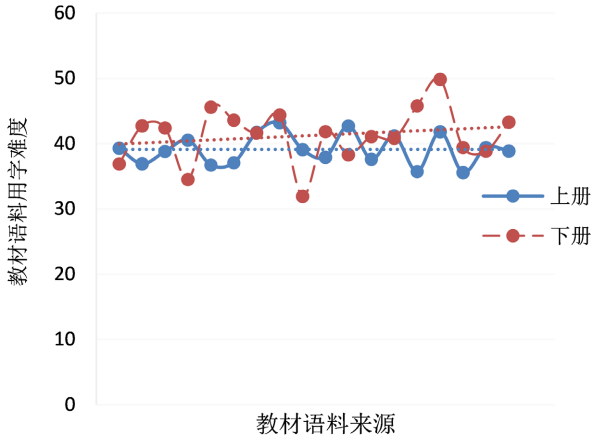

Figure 1. Word difficulty in HSK Standard Course (5)
图 1. 《HSK 标准教程(5)》用字难度

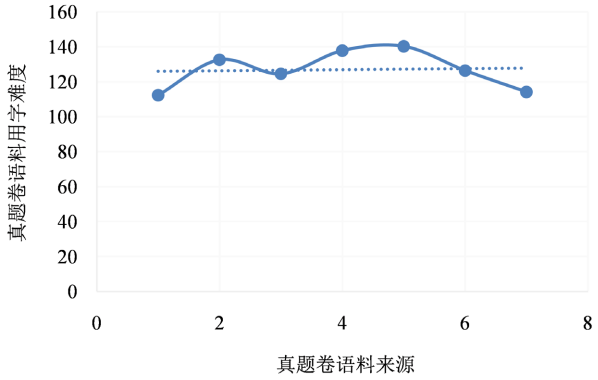

Figure 2. Word difficulty of the new HSK Grade 5 real question paper
图 2. 新 HSK 五级真题卷用字难度

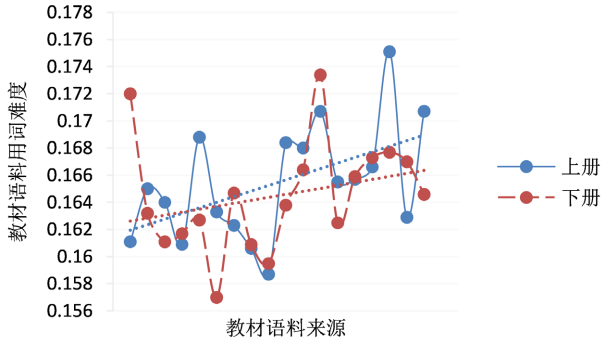

Figure 3. Vocabulary difficulty of HSK Standard Course (5)
图 3. 《HSK 标准教程(5)》用词难度

通过对课文语料的用词难度计算结果能看出，《教程(5)》课文的用词难度虽然未表现出太大的数值差异，但其散点图之间的连线波动远大于用字难度，且与《教程(5)》的用字难度变化趋势不同的是，该课文中的用字难度随时间推移而增加，用词难度却是先缓慢上升，到达某个高度后再缓慢下降，整体呈

“凸”字势。教材的用词难度差距更大的是上册与下册之间而非课文，且上册课文的用词差异大于下册。随着时间推移，上下册间表现出来的难度增长趋势也表现出来较大差异。与课文的用词难度相比，真题卷用词难度表现出持续稳定性。

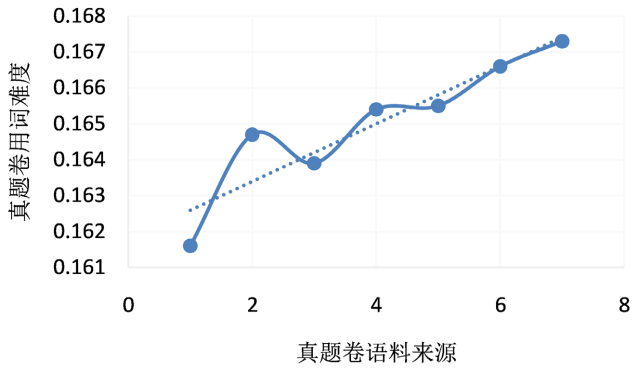

Figure 4. Vocabulary difficulty of the new HSK Grade 5 real test paper
图 4. 新 HSK 五级真题卷用词难度

对比真题卷与课文中的超纲词，《教程(5)》中的超纲词词次比(12.20%)高于真题卷(10.84%)。在用字情况中出现的超纲字，《教程(5)》表现为表示地名的汉字，真题卷表现为出现在历史故事中的汉字。与用字情况中的超纲字有所不同，超纲词则表现为专有名词、数词等。此外，由于真题卷语料大于教材语料，因此真题卷中出现的成语数量也大于教材，但二者的成语复现率都偏低，绝大部分成语仅出现一次。

然后，利用断句标点计算出两个语料的平均句长，并绘制散点图 5 与图 6：

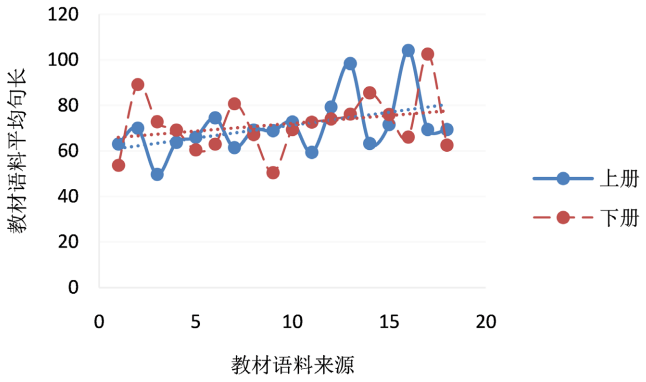

Figure 5. Average sentence length of HSK Standard Course (5)
图 5. 《HSK 标准教程(5)》平均句长

根据上文中的散点图可以看出，《教程(5)》上、下册语料的平均句长存在较大波动。从趋势线来看，两册课文的平均句长都呈上升趋势，《教程(5)》上起点比《教程(5)》下低，但上升趋势快于《教程(5)》下，最终的结束点高于《教程(5)》下，《教程(5)》下的情况与《教程(5)》上相反，这与《教程(5)》上、下册的用词难度表现相同。：在《教程(5)》上的平均句长中，上升趋势较之前的用字、用词来说更为缓慢。《教程(5)》上的单篇语料平均句长大部分都集中在 65~75 这个区间。《教程(5)》下的单篇语料数值大都集中在 65~80 之间，二者的集中区间存在一定差异。

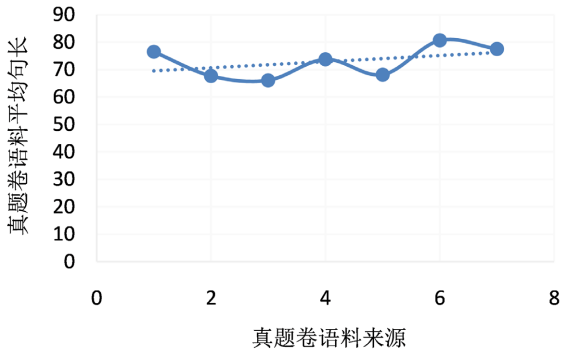

Figure 6. The average sentence length of the new HSK Grade 5 test paper
图 6. 新 HSK 五级真题卷平均句长

与《教程(5)》的单篇语料平均句长相比，真题卷的单篇语料平均句长表现出更稳定的趋势，不同年份的真题卷平均句长有所不同，但篇与篇之间的平均句长差距表现得非常小。在单个语料内，真题卷的平均句长未表现出随时间推移而上升或下降的趋势，这与前两章计算出来的用字、用词变化趋势有所不同。

最后，计算出教材语料与真题卷语料的可读性分数，并绘制散点图图 7 与图 8：

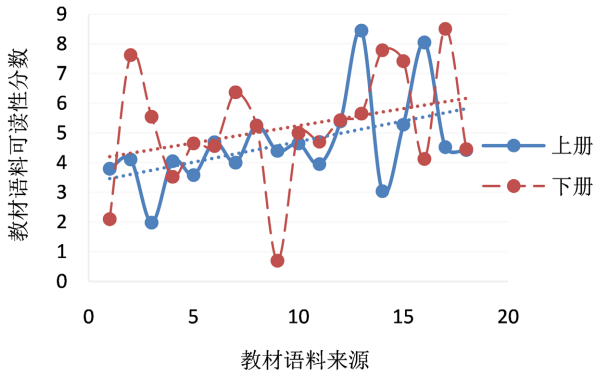

Figure 7. Readability score of HSK Standard Course (5)
图 7. 《HSK 标准教程(5)》可读性分数

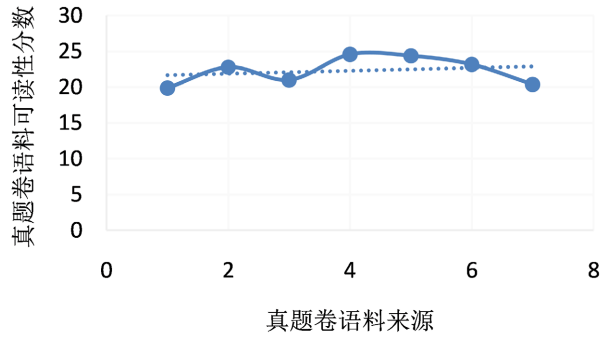

Figure 8. The average sentence length of the new HSK Grade 5 test paper
图 8. 新 HSK 五级真题卷平均句长

从散点图可以清晰地看出，《教程(5)》下的可读性分数趋势线不管是起点还是终点，整体来说都是高于《教程(5)》上的。此外，《教程(5)》下的可读性分数波动性远大于《教程(5)》上。从《教程(5)》中单篇语料的可读性来看，不管是上册还是下册，语料之间的可读性都表现出较大差距。不管是从可读性分数统计表还是散点图来看，《教程(5)》的语料难度都存在较大波动性，除此之外，语料难的可读性分数呈上升趋势，这说明语料的难度也在逐渐上升。

计算出新 HSK 五级阅读部分的可读性分值后，能发现真题卷中单篇语料的可读性分数与前面的影响因素保持一致：单篇语料间的可读性分数相差不大，且绝大部分都集中在 20~25 这个区间内。

从上述计算结果来看，教材与真题卷之间存在着较大差异，将两个语料的整体结果进行计算后，结果见表 2：

Table 2. Overall readability comparison between HSK Standard Course (5) and new HSK Level 5
表 2. 《HSK 标准教程(5)》与新 HSK 五级整体可读性对比

语料来源	平均句长	用字难度	用词难度	可读性分数
《HSK 标准教程(5)》	69.8524	274.661	0.1648	51.1624
新 HSK 五级	70.9683	126.928	0.1649	50.6107

从表格中可以看出，影响《教程(5)》与真题卷的可读性结果的三个因素中，用词难度与平均句长都保持了高度一致，差距最大的是语料的用字难度。即使抽样的真题卷整体语料大于《教程(5)》，但最终的计算结果仍显示《教程(5)》的用字难度大于真题卷，从而导致最终的可读性分数高于真题卷。

以上结果显示，教材与真题卷语料的平均句长相似、平均句长波动幅度都与用字难度的波动相似、平均句长最小值与用字难度最小值保持一致，从平均句长来看，《教程(5)》的整体设计遵循由易到难的规律，呈现出波浪式的缓慢上升，一开始的难度略低于 HSK 五级，随着学习进度的增加，难度也逐渐增加，直至学生完成上、下两册的学习后，最终汉语水平达到甚至略高于考试要求的五级水平。

二者存在以下差异：就单个语料的可读性分数来看，《教程(5)》中的语料可读性具有很大的不稳定性，语料间的可读性分数差异较大；真题卷的单个语料可读性分数高于《教程(5)》，这一特征在语料的用字部分也体现得非常明显；从二者的可读性分数趋势线来看，语料的可读性分数都呈缓慢上升趋势，但真题卷的上升趋势缓于《教程(5)》；《教程(5)》的可读性分数高于真题卷而真题卷的可读性比教材更稳定。

4. 结论与建议

从以上结果更加清晰地了解了教材与真题卷间的差异，并据此为教师、学生、教材编写者提出相应建议。

4.1. 对教师的建议

4.1.1. 注重超纲词教学

从《教程(5)》与真题卷中的超纲字与超纲词来看，二者之间的最大相同点是有许多与中国传统文化相关的字词。超纲字词会增加语料的阅读难度，因此教师在教学中，可以让学生通过看偏旁、联系上下文等方式猜词，增强学生的猜词能力。

4.1.2. 注重对话题的分析与归纳

教师在训练学生的汉语阅读能力时，应注意选取不同话题、不同题材的阅读材料进行训练，以帮助

学生更好地理解不同话题与不同文体间的差异。

4.1.3. 增加学生的阅读数量

由于教材中包含的内容有限，教师可以举办小型读书会、阅读分享会等活动，由教师指导、学生主动参与，选取与真题卷的语料难度相当、话题相似的文本进行阅读，以帮助学生提升阅读汉语文本的兴趣，提升中文水平。

4.2. 对学习者的建议

4.2.1. 自行进行拓展阅读

学习者平时可以多进行拓展阅读。在选择阅读材料的时候，可以有针对性地选择报刊、新闻等与考试话题接近的阅读材料。

4.2.2. 阅读过程中刻意使用阅读技巧

学习者在学习阅读技巧后，更重要的是在实际的阅读过程中加以运用，做到熟能生巧。阅读能分为精读和略读，学生在进行阅读时，应注意针对不同的阅读题型，使用不同的策略。

4.2.3. 善用网络资源进行学习

学生可以利用网络便利，通过不同的网站、学习 APP 等了解与中国相关的时事，在了解中国的发展步伐的同时，提升自己的阅读水平。

4.3. 对教材编写者的建议

4.3.1. 注意课文间的难度差异

在教材的同一册同一单元内、同一单元内的不同课文间，其难度存在较大差异。这可能会导致学生的最终学习效果受到影响。因此编写者在编写课文时，应避免这一情况出现，以便为学生提供更好的客观学习条件。

4.3.2. 对部分课文内容进行修改

教材编写应遵循针对性、实用性、科学性、趣味性、系统性五大原则[16]。对教材中选用的少部分时效性较强的课文，再版时应注意将其进行适当替换，将其替换为近年出现的其他内容。

4.3.3. 与相关网站合作

不管是时代发展还是受新冠疫情影响，越来越多学习者选择主动上网站搜索信息，相关的中文学习网站开发、资源合作共建活动也开展得如火如荼。教材编写者可以在教材开发的基础上，与相关中文学习网站合作，一方面可以为学习者提供多样化的学习平台；另一方面教材也可以保证相关学习语料的及时更新，弥补教材语料滞后这一短板。

基金项目

云南大学第十三届研究生科研创新项目“基于语料库的综合教材对比研究”，项目编号：2021Z038。

参考文献

[1] 全球开考! HSK 考生人数增长显著[EB/OL]. http://www.chinese.cn/page/#/pcpage/article?id=709, 2021-06-09.

[2] 汉语水平考试(HSK) [EB/OL]. http://www.chinese.cn/page/#/pcpage/project?id=113, 2023-04-23.

[3] 郑英, 郑玥, 张家维. HSK 试卷架构对 1-3 级考生成绩的影响——以英语母语者为例[J]. 国际中文教育(中英文), 2021, 6(3): 50-59.

[4] 罗民, 张晋军, 谢欧航, 黄贺臣, 解妮妮, 李亚男. 新汉语水平考试(HSK)质量报告[J]. 中国考试, 2011(10): 3-7.

[5] 吕中舌. 可读性与英语教材[M]. 北京: 世界知识出版社, 2003: 117.

[6] 王蕾. 文本可读性公式研究发展阶段及特点[J]. 语言教学与研究, 2022(2): 29-40.

[7] 周东杰, 郑泽芝. 可读性研究综述[J]. 泉州师范学院学报, 2020, 38(1): 55-63.

[8] 国家汉办/孔子学院总部. 汉语水平考试真题集 2012 版[M]. 北京: 商务印书馆, 2012.

[9] 国家汉办/孔子学院总部. 汉语水平考试真题集 2014 版[M]. 北京: 高等教育出版社, 2014.

[10] 国家汉办/孔子学院总部. 汉语水平考试真题集 2018 版[M]. 北京: 人民教育出版社, 2018.

[11] Che, W.X., Feng, Y.L., Qin, L.B. and Liu, T. (2021) N-LTP: An Open-Source Neural Language Technology Platform for Chinese. *Proceedings of the* 2021 *Conference on Empirical Methods in Natural Language Processing*: *System Demonstrations*, Online and Punta Cana, 7-11 November 2021, 42-49.

[12] 刘华. “汉语助研”语料库建设与统计工具[EB/OL]. http://www.languagetech.cn/corpus/tools.aspx, 2023-04-23.

[13] 刘华. 语料库语言学——理论工具与案例[M]. 北京: 外语教育与研究出版社, 2020: 49.

[14] 郭望皓. 对外汉语文本易读性公式研究[D]: [硕士学位论文]. 上海: 上海交通大学, 2009.

[15] 李昊源. 关于汉语分级阅读可读性公式应用的研究[D]: [硕士学位论文]. 郑州: 郑州大学, 2020.

[16] 刘珣. 对外汉语教育学引论[M]. 北京: 北京语言大学出版社, 2000: 314-316.