汉字属性标注对汉语分级阅读文本过滤参数的投射

盛玉麒
香港城市大学
2016—12—12

一、解题

- 1.1 分级阅读
- 针对不同年龄、阅读水平和理解能力提供不同难度的文本材料；
- 分级阅读的目的
- 提高阅读理解能力，培养阅读习惯
- 有利于情智发育、知识增加、心理发展、综合素质的提升。

1.2文本过滤参数

- 文本分级的原则和标准
- 按年龄或年级分；
- 按阅读水平分级。
- 文本分级涉及到：
- 认知能力
- 知识背景
- 目标期望值
- 文本难易度

1.3国外可参考指标

- 美、英发达国家对分级阅读的研究
- 分级阅读理论；
- 科学分级的原则；
- 能力水平测试的标准；
- 读者能力与文本能读之间的适合度参数。
- 涉及到教育学、心理学、认知科学、社会学等相关学科。
- 政府行为、“总统工程”

- 1997年到2000年凯瑟琳.斯诺带领由18名著名学者组成早期阅读委员会，
- 建立了美国早期阅读系统的理论和儿童阅读的目标体系，
- 报告名为《培养成功的阅读者》。
- 2001年小布什总统通过《不让一个孩子落下》法案(No Child Left Behind Act)，
- 奥巴马政府2009年7月宣布《冲向卓越》（Race to the Top）计划，

1.3英国的分级阅读研究

- 牛津大学花了20年的时间研究分级阅读，
- 1998年教育部一年投入3700万英镑,倡导“英国阅读年”活动.
- 口号是“把阅读进行到底”。

1.4日本的分级阅读

- 上世纪60年代初开展“亲子20分钟阅读”的分级阅读运动；
- 1999年日本国会通过决议，规定2000年为儿童阅读年；
- 借鉴英国全民性的“阅读起跑线”运动，
- 口号是“阅读让希望与梦起飞”。
- 确定每年的4月23号是全国阅读日。

1.5大陆的分级阅读

- 近几年逐渐“热”起来，方兴未艾。
- 不少出版社看到商机，推出系列丛书“桥梁书”、“阶梯阅读”产品，
- 还有研究机构推出相关标准，如南方分级阅读研究中心在分级阅读产品出版、制定自己的分级阅读标准方面取得了一些区域性成果。
-

二、美国分级阅读体系和标准

- 玛丽•克莱，阅读校正体系（Reading Recovery Program）
- 凡塔斯、潘尼，指导阅读体系(Guided Reading)
- 发展性阅读评介体系(DRA)
- 阅读能力等级计划(Degrees of Reading Power <DRP>)

2.1 分级阅读的原则参数

- 1）综合能力训练原则
  - 语言发展、
  - 阅读能力、
  - 写作能力，
- 2）系统性关系原则：
  - 语言、
  - 阅读；
  - 写作；
- 3）循序渐进的原则。

2.2阅读文本的分类标准

- 美国对英语阅读文本的分类：
  - 文学作品，
  - 信息类文章，
  - 基本技能。
- 不同类型的文本有具体的标准
- 有关文学作品的阅读标准，按年级制定出不同的细则，值得参考。
- 例如：

一年级

- 能够提问和回答关于故事中的细节的问题。
- 能够复述故事的主要情节。
- 能够描述故事中的人物和地点。
- 能够判断表达人物感受，心情的词和短语。
- 能够分清谁是叙述者。
- 能够比较故事中不同人物经过的不同事件。

二年级

- 能够提问和回答关于故事中“who，where，when，why，how“（人物，地点，时间，原因，过程）的问题
- 能够详细地复述故事（主题包括来自不同文化的民间小故事，寓言）
- 能够判断故事的中心思想。
- 能够判断诗歌中的押韵，头韵。
- 能够比较同一个故事的不同版本（比如，不同作者的改编的版本）

三年级

- 能够详述来自不同文化的小故事，寓言，传说，神话故事等，并解释故事中的细节，情节是如何体现其中心思想。
- 能够分辨叙述者的观点，故事中人物的观点，以及自己个人观点。
- 能够解释不同的段落，章节在全文的作用。
- 能够比较同一个作者的不同作品中的人物，情节发展等。

四年级

- 能够通过文本中的细节和例子来分析故事，诗歌，剧本的主旨。
- 能够通过文本中的细节和例子来描述故事中的人物，场景和主要情节。
- 能够比较故事的文本和改编的其他表现方式，比如电影，话剧等。
- 能够比较不同的文化的作品对同一个主题的不同的阐述方式。

五年级

- 能够准确地引用文本中的句子来回答问题。
- 能够客观地概括故事内容。
- 能够对比故事的不同人物或场景。
- 能够描述叙述者的观点如何影响了他的叙述方式。
- 能够对比同一种文学类型的作品。

六年级

- 能够准确的引用文本中的句子，并分析句子的意思和前后文的联系。
- 能够分析故事情节或戏剧章节如何展开,递进,结束。
- 能够分析某一句话在文章中的作用。
- 能够分析作者如何建立人物或叙述者的观点。
- 能够表达自己阅读故事，剧本的感受。
- （七—12年级从略）

3.1美国蓝思阅读框架的参数

- 蓝思阅读框架是全美最具公信力的阅读难度分级系统
- 目标：开发一套衡量学生阅读水平和标志文章难易程度的标准，科学衡量：
- 文章难易度；
- 阅读能力。

3.2文章难易度

- 从两个角度确定：
- 1）语义难度
- 2）句法复杂性

3.2.1语义难度

- 以词义理解为主。包括以下参数：
- 1）常用词占比
- 2）生词量占比

3.2.2句法复杂程度

- 主要看句子长度
- 长句的占比

3.3分级阅读标准参数的细化

- 3.1文本难度相关因素：
- 1）书本的规模；
- 2）页面表现形式；
- 3）插图信息量；
- 4）概念的复杂性；
- 5）单词复现量；
- 6）内容意义的可预测性。

3.4阅读水平单词测试法

- 判断原则：
- 根据对孤立单词认读量占比判断水平等级。
- 孤立单词的认读理解是否需要上下文语境信息提示；
- 通过抽样语料库统计分析，制作出分级词汇常模量表。

Preprimer	幼儿园	1 年级	2 年级	3 年级
see	you	road	our	city
play	come	live	please	middle
Me	not	thank	myself	moment
At	with	when	town	frightened
run	jump	bigger	early	exclaimed
Go	help	how	send	several
and	Is	always	wide	lonely
look	work	night	believe	drew
can	are	spring	quietly	since

4 年级	5 年级	6 年级	7 年级
decided	scanty	bridge	amber
served	business	commercial	dominion
amazed	develop	abolish	sundry
silent	considered	trucker	capillary
wrecked	discussed	apparatus	impetuous
improved	behaved	elementary	blight
certainly	splendid	comment	wrest
entered	acquainted	necessity	enumerate
realized	escaped	gallery	daunted
interrupted	grim	relativity	condescend

8年级	9 年级	10 年级	11 年级
capacious	conscientious	zany	galore
limitation	isolation	jerkin	rotunda
pretext	molecule	nausea	capitalism
intrigue	ritual	gratuitous	prevaricate
delusion	momentous	linear	visible
immaculate	vulnerable	inept	exonerate
ascent	kinship	legality	superannuate
acrid	conservatism	aspen	luxuriate
binocular	jaunty	amnesty	piebald
embankment	inventive	barometer	crunch

3.5 阅读速度标准分级参数

- 美国阅读标准给出的每分钟阅读单词量分级参数为：
- 初级： 80 — 158个单词；
- 中级：175 — 204个单词；
- 高级：214—1250个单词。
-

3.6中国儿童青少年分级阅读水平评价标准

- 南方分级阅读研究中心制定
- 把1—9年级分四学段。
- 第一学段（1—2年级）
- 第二学段（3—4年级）
- 第三学段（5—6年级）
- 第四学段（7—9年级）
- 用描述方式，介绍该学段阅读文本选择的原则性建议。
- 缺乏严谨的科学性和可操作性。

第一学段（3—4年级）

- 1.选择内容丰富、形象具体、文字少、故事趣味性强的童话图画书（一年级加注拼音），图画书与文字书所占比例不少于1/2。逐步增加文字的阅读量，让儿童青少年在有趣的图象和文字的结合中，感受阅读的乐趣。
- 2.选择具有更多现实性、体验性、思考性的童话故事、寓言故事、童谣等，使儿童青少年的情趣更加浓厚，吸引其独立阅读完一本书。
- 3.选择带有具体感知的动植物知识的启蒙读物，激励儿童青少年产生更多的科学兴

第二学段（3—4年级）

- 1.选择浅显的具有哲理性的故事，帮助儿童青少年区别现实与幻想的差异，分辨美丑是非善恶，初步认识人类社会。
- 2.增加散文、诗歌、科幻等多种体裁的读物，提供轻松幽默且品味高的作品，满足儿童青少年日益增长的求知欲和阅读的需求。
- 3.增加科普知识，扩大儿童青少年的视野。
- 4.选择具有爱国主义和集体主义精神、具有传统文化精髓的故事，激发儿童青少年的爱国主义情怀。

3.7国内推荐书目举例

- 推荐书目：0—3岁 40种；
- 推荐书目：4—6岁 40种；
- 推荐书目：7—8岁 40种；
- 推荐书目：9—10岁 40种；
- 推荐书目：11—12岁 40种；
- 由于缺乏科学的分级阅读标准评测参数，难免见仁见智。

四、书面汉语的特点与美国分级阅读标准的汉化

4.1 书面汉语的特点

- 1）汉字载体“中介”
- “形、音、义、用”多重属性；
- 2）词语理解
- 推敲字理、考究字义
- 新词创造的“命名学”传统：专有名词、术语、略语、熟语
- 3）句子使用与理解
- 讲究字斟句酌、遣词造句、
- 4）文章分析理解
- 轻句法、重微言大义

4.2美国标准“汉化”的思考

- 1）两个指标：语义难度和句法复杂性
- 2）中文的表现：词义表达透明度
- 3）主要难点：语用原则和语义理解。
- 4）对策思路：
- a 基于“组字成词、组词成句”原理，
- b 将多元复杂性问题转化为基于语料库的汉字属性挖掘，
- c 建立“根词语素相关性构式语块”分级常模量表，作为评估语义难度和句法复杂性程度的“转移矩阵”。

4.3中文分级阅读可参考标准

- 1）现代汉语常用字表（3500字）
- 常用字2500字；次常用字1000字
- 2）现代汉语常用词表（65000词）
- 3）老HSK汉字词汇大纲（8822词、2906字）
- 4）新HSK词汇大纲（5000词、2628字）
- 5）中小学语文教学大纲和分级生词表

4.4“常模量表”的基础

- 性质：评估不同阅读理解能力平均水平的字、词、语的等级量表；
- 具有标准化指标系统的科学性和权威性。
- 功能：衡量评估语言知识程度等级的均衡概率指标。
- 研制：基于大规模平衡语料库的字频词频统计结果、进行抽样检验信度测试和试行、
- 经有关职能部门审核批准可获“标准”颁行。

4.5常模量表研制的语料库建设

- 1）中小学语文教材分级语料库；
- 2）中小学课外阅读抽样语料库；
- 3）中小学作文抽样语料库；
- 4）中小学分科分级教材抽样语料库；
- 5）对外汉语教材抽样语料库；
- 6）当代汉语流通语料库（9类、1400万字）

五、汉字等级属性

- 5.1已有属性标准
- 1）现代汉语常用2500字；
- 2）现代汉语次常用1000字；
- 3）老HSK（8822词）4级共2906字
- 4）新HSK（5000词）6级共2628字

5.2平衡语料库字频统计

- 当代汉语流通语料库
- 《红楼梦》全文语料库
- 《三国演义》语料库
- 《莫言全集》语料库
- 对外汉语教材抽样语料库
- 经典诵读语料库：《论语》、《三字经》、《百家姓》、《千字文》、《弟子规》

流通语料库分类词频统计结果

序号	语料类型	词种	词次
01	经济	31,842	746,363
02	政治	22,727	664,712
03	文化	28,794	346,067
04	卫生	16,437	242,129
05	体育	32,392	839,883
06	法律	15,153	572,392
07	文学	83,724	5,560,765
08	网文	55,101	3,526,128
09	博客	47,894	1,089,197

“当代汉语流通语料库”库结构

山东大学中文信息研究所 盛玉麒

类型	字段名	类型	长度	说明	记录(词次)	词种
带属性标志	CT	C	20	考虑兼类或同形词		
不带属性	CT1	C	20	只考虑词形		
词长	CC	N	1	1 字节		
属性	SX	C	3	三个西文字母		
频数	NUM	N	7	7 字节		
频度	PD	N	8	小数点后保留 5 位		
经济	JJ	N	7	7 字节	746,363	31,842
政治	ZZ	N	7	7 字节	664,712	22,727
法律	FL	N	7	7 字节	346,067	28,794
文化	WH	N	7	7 字节	242,129	16,437
卫生	WS	N	7	7 字节	839,883	32,392
体育	TY	N	7	7 字节	572,392	15,153
文学	WX	N	7	7 字节	5,560,765	83,724
网络	WL	N	7	7 字节	3,526,128	55,101
博客	BK	N	7	7 字节	1,089,197	47,894
分布	FB	N	2	2 字节		
合计			124		13,,587,636	334,064

弟子规

弟子规	字种	字次	流通字次	流通字频
老 HSK1 级	249	667	7603751	43.43
老 HSK2 级	112	202	574982	3.28
老 HSK3 级	42	71	97850	0.56
老 HSK4 级	42	118	35569	0.20
集外字	37	41	8719	0.05
合计	482	1099	8320871	47.52
新 HSK1 级	66	224	4300878	24.56
新 HSK2 级	64	173	1589984	9.08
新 HSK3 级	78	177	1338356	7.64
新 HSK4 级	90	190	706842	4.04
新 HSK5 级	78	144	298452	1.70
新 HSK6 级	62	139	76167	0.43
集外字	44	52	10192	0.06
合计	482	1099	8320871	47.51

单字	字次	字频	老HSK	新HSK	常用	笔画数	流通次	流通频	经济	政治	法律	文化	卫生	体育	文学	网络	博客	分布
勿	43	3.9126	4	f	a	4	161	0.0009	4	3	0	25	3	2	56	38	24	8
人	30	2.7298	1	a	a	2	171768	0.981	4928	8396	21282	4369	3487	7314	55148	36616	14768	9
不	25	2.2748	1	a	a	4	208613	1.1914	6179	4822	4710	3899	2411	9456	70480	66237	18971	9
有	20	1.8198	1	a	a	6	156553	0.8941	8254	6365	7119	4208	2532	8565	52251	39130	14449	9
亲	14	1.2739	1	d	a	9	11944	0.0682	54	59	223	126	66	284	6258	2197	780	9
长	12	1.0919	1	b	a	4	37370	0.2134	2720	1486	1471	529	667	923	10902	15137	1656	9
事	12	1.0919	1	b	a	8	44678	0.2552	1518	2489	5584	818	958	1309	13202	11446	3539	9
为	10	0.9099	1	b	a	4	76996	0.4397	6736	4867	3336	3736	2210	5493	24559	12285	6496	9
无	9	0.8189	2	d	a	4	29611	0.1691	852	624	858	935	256	1810	12255	6858	2612	9
心	9	0.8189	1	c	a	4	48847	0.279	992	958	149	863	882	1962	18090	17843	3589	9
则	9	0.8189	2	d	a	6	8029	0.0459	910	848	838	517	183	729	1546	1577	513	9
者	9	0.8189	1	c	a	8	32838	0.1875	2664	1434	6774	1090	1180	1307	11377	1910	2188	9
必	8	0.7279	1	c	a	5	8634	0.0493	564	1123	734	437	445	442	1954	1223	743	9
过	8	0.7279	1	b	a	6	54321	0.3102	3103	1951	1351	790	769	3273	19210	14251	4581	9
即	8	0.7279	2	e	a	7	6667	0.0381	570	279	266	326	192	420	2404	1103	538	9
如	8	0.7279	1	c	a	6	33248	0.1899	1591	867	538	1003	562	1971	11620	8438	3589	9
非	7	0.6369	1	b	a	8	14688	0.0839	758	563	562	354	164	1277	6960	2227	1181	9
轻	7	0.6369	1	c	a	9	10542	0.0602	121	139	174	85	75	662	3939	3919	558	9
问	7	0.6369	1	b	a	6	21604	0.1234	1216	1756	606	500	663	947	6855	5204	1978	9
言	7	0.6369	1	d	a	7	8561	0.0489	346	222	224	452	109	413	3330	1614	1152	9
自	7	0.6369	1	b	a	6	62686	0.358	1566	2164	2566	1660	711	3235	18357	21456	5150	9
道	6	0.546	1	b	a	12	46378	0.2649	720	947	586	897	418	1314	13718	21168	3052	9

《红楼梦》用字频统计降频例样表

序号	单字	字次	字频
1	了	21275	2.52236
2	的	15764	1.86898
3	不	15067	1.78634
4	一	12206	1.44714
5	来	11471	1.36
6	道	11067	1.3121
7	人	10546	1.25033
8	是	10166	1.20528
9	说	9702	1.15027
10	我	9192	1.0898
11	这	7846	0.93022
12	他	7750	0.91884
13	你	7157	0.84853
14	去	6203	0.73543
15	着	6186	0.73341
16	也	6130	0.72677

序号	单字	字次	字频
17	儿	6096	0.72274
18	玉	6056	0.718
19	有	6010	0.71255
20	宝	5812	0.68907
21	个	5723	0.67852
22	子	5486	0.65042
23	又	5213	0.61805
24	贾	5192	0.61556
25	里	5156	0.6113
26	那	4923	0.58367
27	们	4902	0.58118
28	见	4806	0.5698
29	只	4686	0.55557
30	太	4313	0.51135
31	便	4075	0.48313
32	好	4054	0.48064

《红楼梦》用字功能属性等级分布统计

红楼梦	字种	字次	字频	流通字次	流通字频
老 HSK1 级	780	572808	67.91	13516666	77.19
老 HSK2 级	780	85841	10.18	2516191	14.37
老 HSK3 级	547	26752	3.17	675439	3.86
老 HSK4 级	600	18600	2.20	334440	1.91
集外字	1759	138454	16.40	182001	1.04
合计	4466	842455	99.86	17224737	98.37
新 HSK1 级	168	327478	38.83	6754624	38.58
新 HSK2 级	172	123766	14.67	2824276	16.13
新 HSK3 级	273	85283	10.11	2711353	15.49
新 HSK4 级	436	65960	7.82	2423378	13.84
新 HSK5 级	611	60799	7.21	1542028	8.81
新 HSK6 级	811	30621	3.63	665331	3.80
集外字	1995	148548	17.60	303747	1.73
合计	4466	842455	99.87	17224737	98.38

对外汉语教材抽样语料库汉字等级分布统计表

汉语教材	字种	字次	字频	流通字次	流通字频
老 HSK1 级	781	113178	43.14	13612272	77.74
老 HSK2 级	686	16230	6.10	2434177	13.90
老 HSK3 级	329	2533	0.96	541032	3.09
老 HSK4 级	204	1373	0.52	160796	0.92
集外字	92	373	0.14	29342	0.17
合计	2092	133687	50.86	16777619	95.82
新 HSK1 级	172	61695	23.52	6832724	39.02
新 HSK2 级	174	22841	8.71	2835562	16.19
新 HSK3 级	264	21796	8.31	2705592	15.45
新 HSK4 级	433	16052	6.12	2439658	13.93
新 HSK5 级	510	7941	3.03	1449872	0.28
新 HSK6 级	386	2535	0.96	426617	2.44
集外字	153	827	0.31	87594	0.50
合计	2092	133687	50.96	16777619	87.81
常用字 a	1903	132705	50.58	16687705	95.31
次常用字 b	140	797	0.30	76443	0.44

问题与讨论

- 1）分级阅读文本过滤标准中的语用参数是一个复杂集；
- 2）汉字功能属性对参数集的投射具有不完备性；
- 3）基于大规模平衡语料库的构式语块常模量表研究，具有重要的应用价值。
- 4）中文分级阅读是一个宏大的社会性系统工程。需要多专业协同攻关，以望克成。
- 本研究不当之处，敬请不吝赐教。

参考文献

- 1当代汉语流通语料库（山东大学中文信息研究所）
- 2 现代汉语常用字表（国家语委）；
- 3汉语水平考试汉字词汇大纲（老HSK）（汉办）；
- 4汉语水平考试词汇大纲（新HSK）（汉办）；
- 5《红楼梦》附码语料库（山大）；
- 6《三国演义》附码语料库（山大）；
- 7《莫言全集》附码语料库（山大）；
- 8对外汉语教材抽样语料库（山大）；
- 9中国儿童青少年分级阅读内容选择标准（南方分级阅读研究中心）
- 10中国儿童青少年分级阅读水平评价标准（南方分级阅读研究中心）
- 11百度网站资源，恕不尽列，一并致谢。

- 谢谢！

- 盛玉麒 鞠躬
- 2016-12-12
- 13031731940
- Yuqi-sheng@163.com