维普资讯 http://www.cqvip.com

计算机科学 1999Vol 26№ 11

中文阅读难度模型及易读性公式探索

On the Difficulty Model of Chinese Text Reading and Readability Formula

陈阿林
（重庆师范学院计算机中心 重庆 400047）

张 素
（西南师范大学计算机科学系 重庆 400715）

Abstract a lot of work has been done abroad on easy-reading test and dimensional formula of English. In this paper, research work abroad is introduced. On the basis of mathematical methods, statistics, induction and analysis are conducted on Chinese text. A kind of quantitative approach is suggested for Chinese reading difficulty, furthermore, preliminary formula is provided for easy-reading of Chinese text, and verification is performed.

Keywords Reading difficulty, Chinese, Easy-reading formula, Computational linguistics

1. 前言

计算机应用已从数据处理开始迈入知识处理、语言理解阶段，人们对计算机的智能提出了新的要求，机器的智能化则对语言文字的处理深度和广度越来越高。

利用计算机对汉语基础研究在字、词用语方面取得了相当的成果并进入到对汉语词汇属性、句子的分析，文本、语义等处理。由于因特网上大量以 HTML 文本出现的信息数据，使得对文本的处理、加工得到高度重视，文本的阅读难度描述和其定量表达是文本处理的一个方面，国外在计算机出现之前便已开展了研究，所得结果用于许多方面。本文作者以汉语文本语料统计为基础，提出汉语文本阅读难度的模型结构，通过数学统计进行定量化处理，作为初步研究，给出汉语文本易读性的一个公式并进行简易的验证。

2. 英文阅读难度研究概述

英文文本的阅读难度即：易读性测试和量纲公式，国外已开展了几十年，取得了许多成果[2]。从事英文文本易读性测试和量纲公式较早的是 Irving Lorge，他在 1939 年开始研究儿童阅读材料的阅读难度并发表了第一个定量表达公式，Flesch，Kearl，Dale，Chall 等人都对英文文本易读性测试和量纲公式作出了许多贡献。目前使用最广泛的是 Rudolf Flesch 在 1948 年给出的一般成年人阅读材料的 Flesch 公式（按分级）：

$$RE = 206.835 - 0.846wl - 1.015sl$$

$$HI(\text{Human Interst}) = 3.635pw + 0.314ps$$

其中：wl＝每 100 个单词音节标记数；sl＝段单词平均数；pw＝每 100 个词个人的词汇平均值；ps＝每 100 句子个人的句子标记。

Flesch 公式划分阅读材料为 0—100 级，0 级为无法阅读，100 为极易阅读。该易读性公式，成为在英文文本易读度量的首选，用它对美国的三种著名的文献资料，时事周刊-Time、法律专业杂志 Harvard Law Review、保险公司的文件条款进行了计算，Time 的平均易读性指数为 52、Harvard Law Review 平均易读性指数为 32、保险公司的文件条款平均易读性指数为 40—50。

另外，美国国防部为了军队教育，修改了 Flesch 公式，制订 Flesch-Kincaid Grade Level 指数（FKGL），使易读性量的计算仅由英文的单词和句子的平均字符长度来决定：

$$FKGL = 39Xs + 11.8Xwa - 15.59$$

其中 Xwa：平均单词的字母长度

当阅读材料的计算值为 8.0 分值时，表明八年级的学生可阅读和理解该篇阅读材料，标准的写作文章的得分应为 7—8 级，6—10 级之间为较好。

Microsoft 公司在 1994 年推出 Microsoft Word6.0（包括后续版本）时，已将英文易读性计算功能嵌入其中，由此可见文本易读性在文字处理中是有重要作用的。

值得说明的是，英文文本的阅读难度易读性测试和量纲公式的研究均以 1925 年 McCall 和 Crabbs 从大量英文文本阅读材料中提出的一组阅读测试课文为标准，遗憾的是我国汉语专家至今还未提出这类标准。

目前，国内从事英语教学、研究的学者也有采用 Flesch 易读性公式用于英语专业语言教学阅读材料、

维普资讯 http://www.cqvip.com

词汇量选取的选择[3]。

3. 中文阅读难度分析及简化模型

3.1 字、词、句对难度的影响

照"文章由段落构成，段由句子构成，句由单词按相应的语法结构构成"这一通常看法，作者认为中文的阅读难度以句和词为主体。

客观地说，中文阅读难度与中文字、词、句子、语法、文本内容和长度都有直接关系，也与读者的心理，受教育程度，个人专业，兴趣爱好等相关。句子包含着丰富的语法结构，对阅读难度有直接关系，语法就语义的表达来看，由于阅读文本传达信息的对象是具有智能的人，所以语法结构对阅读理解困难度量纲显著性应该不太大，这是作者考虑仅使用句子类型，不参考具体语法，将句型作为难度基本计算量纲的依据。

在人的主观世界中词代表一个抽象的基本概念，不同的单词表示着不同的概念，被看作客观世界的独立事物或过程，它在文本中占有重要地位。加拿大语言专家麦克奴认为，一旦课文、章篇中的生词超过 30%，文章将无法阅读，一般人的看法亦同，我认为将词作为文本阅读难度的基本计算量纲之一是学术界认可的。

汉字，虽字集巨大，但常用字并不多，按国内字频统计的结果和香港安子介先生的研究和观点，能识 2000 汉字便能读懂文章的 97.4%[2]，而高频字大多数又为高频单字词，故对阅读难度的因素作者将其集中于句、词。

3.2 用语定义

为了本文用语的准确性，以下给出本文用语的描述性规定：

汉字：符合 GB2312—80 的符号；

汉字词（也简称词、词汇）：由汉字所构成的符号串；

句子：由汉字词按一定语法结构关联构成，并以标点结尾。标点是逗号、句号、分号、冒号、问号、惊叹号等，语法结构符合现代汉语的描述即主谓宾、定状补；

句型：句子的基础结构类型，是形式上具有相同组合关系句子的抽象与概括；

属性：对句子或汉字词汇特征的一级描述值，是一个有限集合，并可量化。

文本：由多个意义上互相关联，独白式的句子组成的语言单位。

3.3 模型结构形式

按字、词、句、句子类型构成文章的概念，可以将中文易读难度（CRE）归纳成下列表达形式：

$$CRE(T)=f(c,w,s,st) \tag{1}$$

其中，T 代表中文文本，c 为 T 构成的文本集的汉字计数，w 为 T 构成的文本集的词计数（区分重复计数和非重复计数），s 及 st 则代表 T 中句子计数和句型计数。

4. 模型的精确化的技术处理

4.1 属性标注

要完成模型的精确化，公式（1）必须是定量化的表达形式，对文本的分词、词类处理、句型句类处理和相关属性都应仔细考虑。属性标注是对文本的词、句型的标注，以得到统计上的定量计量。

单词的属性标注采用了计算机标注加人工校对修改。首先是机器标注，它以属性词表为依据，但由于歧义和汉语兼类词多，故必须进行人工校对。词汇属性则按文[4]中对汉语词汇的分类，共分 13 类（名、动、形容、副、助、代、连、叹、象声、语气、数、量、介），另外，为考虑文本用词与文本阅读难度的关系，特别是为量化上的要求和方便，在处理词属性时加入了现代汉语频率词典[2]中词级"使用度"（Usage）。文[3]中共 8548 词条，按使用频率"使用度"共分为 1—635 级，作者文本处理的词汇库超过了 4 万，所以，当不在此 635 级别的词汇，作者定义其使用度为 700。

句型标注采用手工方式，标注标准采用了清华大学中文系罗振声先生等人承担的国家自然科学基金项目[6]中的结论，即汉语句型频度表，共分句型为 191 类（原分为 209 种[5]）。

4.2 汉语词、句分布频率对难度计量上的对策

汉语词、句用语的分布具有不均匀性，文[6]汉语句型频度表对汉语句型频度分级为高频、次频、低频以及生僻等四级句型，各占累计频率比例为 87%，11%，1.9%，0.1%，（不同专业领域内的文本句型分布可能有少量差别）；文[2]汉语频率词典的前 1000 高频词使用累计频率占 73.13%，使用度为 1—15 的词汇（的、了、是、一、不、我、在、有、他、个、这、着、就、你、和）累计频率比例已达到 21.3297%。作者认为不属高频率使用的句、词对阅读困难有明显的影响，为了阅读难度量纲上的分析，减低高频句、词在统计计量上的突出分值，我们对句型分布频率和词汇使用度值按公式（2）、公式（3）进行了作者称之为的"规一化"处理（作用为提升非高频率用句、词的计量值）。

$$S_1=\beta_s/(s_f+\alpha_s) \tag{2}$$

$$W_1=\beta_w/(w_f+\alpha_w) \tag{3}$$

其中：S_1，W_1 是句、词级使用度规一化的值，s_f、w_f 是原句频和词级使用度，α_s、α_w 和 β_s、β_w 是作者使用的规一化参数。

4.3 公式的精确化

根据模型公式（1）和大量语料统计，通过多元回归

维普资讯 http://www.cqvip.com

和迭代的数学处理，作者导出一组初步的中文阅读难度计算公式。其中表达公式(1)的 s、t 参数是通过使用平均句，词长(Lsa、Lwa)来反映的，另外还考虑了词、句分布频率对难度计量的影响($Xw\Sigma$、$Xs\Sigma$)。

$$CRE(T)=f(c,w,Lsa,Lwa,Xw\Sigma,Xs\Sigma)=$$
$$109.411-0.08*C-0.11*W+5.57*Lsa-$$
$$4.787*Lwa+0.275*Xw\Sigma-0.483*Xs\Sigma+\Gamma \quad (4)$$

其中 C：文本汉字计数；W：文本词计数

Lsa：平均句字长；Lwa：平均句词长

$Xw\Sigma$：文本中全体词集使用度规一化值之和，即：$\Sigma S1$，见公式(2)

$Xs\Sigma$：文本中全体句子使用频度规一化值之和，即：$\Sigma W1$，见公式(3)

参数 Γ 是为考虑文体取材，如写作题材、风格等，加入的校验因子，取值将进一步讨论。

由于没有中文的标准阅读材料，我们选取了 20 篇中文文本，采用了多人阅读打分进行等级划分，并取打分取均值的方式，共分 5—6 个等级，1 级为最简单，级别增高，难度增加。验证文本(编号)列表如下：

1 《哈里》，英语趣味小品一册 P13，中南工大出版社
2 《雷德》，同上第二册 P106
3 《斯科特太太》，同上第二册 P55
4 《冬天已经来临》，同上第一册 P83
5 城堡，小学生组词造句手册，中国国际广播出版社
6 《英雄》，英趣味小品五册 P100，中南工大出版社
7 骆驼和羊，小学课本第三册，人教出版社
8 我的外公，全国小学生优秀作文精选，大连出版社
9 海底世界，小学课本第七册，人教出版社
10 聪明的仆人，英趣味小品五 P90，中南工大出版社
11 十里长街送总理，小学课本第九册，人教出版社
12 李时珍，小学课本第七册，人教出版社
13 鹤子姑娘，小学生作文向导 1997 年 7—8 期
14 两个铁球同时着地，小学课本八册，人教出版社
15 扫雪，中学生作文选
16 为人民服务，毛泽东选集
17 山的那一边，中学教科书语文二册，人教出版社
18 金色花，学语文杂志，1998 年初中 10 月号
19 妈妈请不要太爱我，小学生作文向导，1996 7—8 期
20 谈毅力，中国中学生作文大全，上海远东出版社

令 β_i、β_w 为 1，取 α_i 为 1 和 10、α_w 为 100 和 10，按式(4)计算得到结果如表 1。

这里，20 个课文人为划分如下(五级)：

1 级为{1,2,3,4,5}；2 级为{6,7,8}；3 级为{9,10,11,12}；4 级为{13,14,15,16}；5 级为{17,18,19,20}。

如图 1 所示，将计算的数值分布仍按 5 个等级划分，当取 $\alpha_w=10$ 可划分为下：

1 级{3,2,5,4,1}；2 级{8,7,6}；3 级{11,9,10,12}；4 级{13,14,15,16}；5 级{18,20,17,19}。将计算的数值分布仍按 5 个等级划分时全部符合人为划分的等级，或按 6 个等级则第 20 文由于具有一定的引申意义，其排列应属第 6 类，但图上看出应在五类。如取 $\alpha_w=100$，将与人为划分有一定的差别，但从大分类上是可以接受的，可见公式还应进一步精确化，α、β 值也可进行调整。

表 1

课文编号	$\alpha_w=100$ $\alpha_i=1$	$\alpha_w=10$ $\alpha_i=1$	$\alpha_w=100$ $\alpha_i=10$	$\alpha_w=10$ $\alpha_i=10$
1	92.58	101.25	93.71	102.38
2	102.65	108.23	103.63	109.20
3	106.29	111.34	107.44	112.50
4	99.22	103.04	100.48	104.31
5	95.28	104.45	96.96	106.13
6	69.44	84.72	71.76	87.05
7	81.49	89.50	83.05	91.06
8	80.94	91.20	83.06	93.32
9	61.09	76.06	64.36	79.33
10	54.89	73.18	56.87	75.16
11	64.54	76.27	66.77	78.51
12	56.23	70.83	59.95	74.56
13	38.83	57.47	42.00	60.63
14	34.38	56.70	38.28	60.59
15	37.85	55.92	42.52	60.59
16	31.45	55.01	35.62	59.18
17	−17.39	9.36	−11.38	15.36
18	8.67	32.77	13.58	37.68
19	−23.85	7.76	−17.58	14.03
20	−2.72	22.70	4.98	30.39

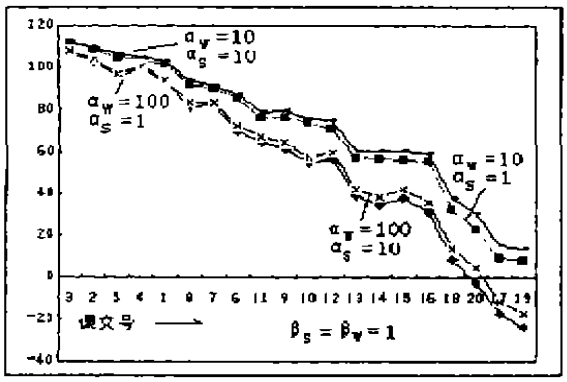

图 1 计算 CER 值图示

(下转第 27 页)

维普资讯 http://www.cqvip.com

概率有关，通常可使用 0,1 等概率的随机序列。

3. 获胜概率的计算

各码字的优先程度用它们相应的获胜概率来度量。码字在随机序列中的获胜情况可用状态转换图来描述。例如，对码字 01 与 11 获胜情况可用图 1 所示状态转换图表示。

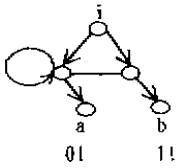

图 1

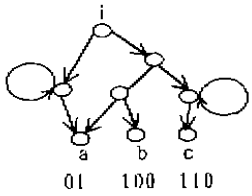

图 2

从根节点 i 处开始，细线表示输入为 0，粗线表示输入为 1。a 节点表示 01 码字获胜，b 节点表示 11 码字获胜。每条支路的传输系数等于概率 1/2。由梅森公式[3]可得节点 a(01)的获胜概率为：

$$Pa=(1/2*1/2+1/2*1/2*1/2)/(1-1/2)=3/4$$

节点 b(11)的获胜概率为：

$$Pb=(1/2*1/2)/(1-1/2)/(1-1/2)=1/4$$

计算机模拟结果：产生长为 10000 个(0,1)随机序列，用上述仲裁方式可得 01 获胜 2543 次而 11 获胜 838 次。与理论计算相符。又如，若取 a 为 01，b 为 100，c 为 110，状态转换图如图 2 所示，用梅森公式计算可得：

$$Pa=[1/2*1/2(1-1/2)+1/2*1/2*1/2(1-1/2-1/2+1/2*1/2)]/[1-1/2-1/2+1/2*1/2)]=5/8$$

类似可得：

$$Pb=1/2*1/2*1/2(1-1/2-1/2+1/2*1/2)/(1-1/2-1/2+1/2*1/2)=1/8$$

$$Pc=1/2*1/2*1/2(1-1/2)/(1-1/2-1/2+1/2*1/2)=1/4=2/8$$

类似地用计算机模拟，对 10000 个 0,1 随机序列，01，100 和 110 的获胜次数分别为 1922,755,380。与理论计算相符。其他如取 011 与 100 则获胜概率为 1/2,1/2，对应模拟获胜次数为 998,996。而码字 011,010,110,111 的获胜概率分别为 3/8,3/8,1/8,1/8，在 10000 个 0,1 随机序列的模拟中获胜次数分别为 931,949,296,294。若选用码字 010,011,100,101，获胜概率分别为 1/4,1/4,1/4,1/4。用 10000 个随机 0,1 序列模拟所得获胜次数分别为 617,634,635,628。

结论 用状态转换图能很好描述随机争用仲裁方式中各指定识别码在 0,1 随机序列中的获胜状况，并可用来计算各识别码的获胜概率。只要改变各设备的识别码模式就可方便地实行公平竞争或优先竞争。调节 0,1 随机序列中的 0,1 出现的概率也可用来调节获胜概率。

参 考 文 献

1 Stanbaum A S. Computer Network(3rd Ed). 1998(中译本)

2 胡道元编．计算机局域网．清华大学出版社，1996

3 赵永昌编．信号流图及其应用．人民邮电出版社，1975

(上接第 44 页)

结束语 作者认为，对中文阅读难度定量化的研究和精确性提高以及方法的拓广，有着潜在的应用前景，如对机器翻译的辅助量纲，Internet 中文信息搜索的一个指标量，变通使用可作为专业查询的背景或智能理解背景的某些指标等，本文给出的公式有待专家的指正和进一步探索，我们正在采用扩大统计语料规模、数量，拓宽分析面，采用多种数学工具并对 Γ、α、β 因子按多种取值寻找对语料文本的相关因素，进一步进行研究，以期获得更佳结果。

参 考 文 献

1 Chapman L J, Czerniewska P, Eds. Reading from Process to Practice. Lndon and Henley Routledge & Kegan Paul in association with The Open University Press ISBN 0 7100 0055 3

2 计算机时代的汉语和汉字研究．清华大学出版社(研讨会文集)，1996

3 现代汉语频率词典．北京语言学院出版社，1986

4 信息处理用现代汉语分词规范及自动分词方法．清华大学出版社．广西科学技术出版社，1994

5 罗振声，郑碧霞．汉语句型自动分析和分布统计算法与策略的研究．中文信息学报，1994.2

6 罗振声，郑碧霞，孙长健．汉语句型频度统计的研究 同 2

7 朱学锋，俞士汶．自然语言处理与语言知识库．同 2

8 陈立为，袁琦主编．计算语言学进展与应用．清华大学出版社，1995

9 张月杰，姚天顺．基于特征相关性的汉语文本自动分类模型的研究．小型微型计算机系统，1998.8

10 吴念，张荣建．英语速读技巧．重庆出版社，1995