AlphaReadabilityChinese:汉语文本可读性工具开发与应用*

雷　蕾[1]　韦瑶瑜[2]　刘康龙[3]

(1. 上海外国语大学语料库研究院　上海　201620;
2. 上海外国语大学国际教育学院　上海　201620;
3. 香港理工大学中文及双语学系　香港　999077)

摘　要:文本可读性是文本的重要语言特征,已被广泛应用于多个学科的研究。现有可读性研究大多聚焦于英语文本可读性指标和工具的开发,而针对汉语的研究尚处于起步阶段。另外,相关研究多采用词汇和句法等表层特征,且主要聚焦于国际中文教育教材和学习者文本展开。本文旨在报告我们开发的汉语文本可读性工具AlphaReadabilityChinese。该工具包括词汇、句法、语义三个维度共九个语言指标,采用了更成熟、稳健的算法开发,是通用的汉语可读性指标和工具。本文以金庸和古龙两位作家的作品为语料测试该工具的有效性。结果显示,两位作家的作品在可读性指标上具有较大差异,金庸作品文本的可读性显著弱于古龙作品。测试结果同时表明,该工具所包含的可读性指标可以很好地区分两位作家的作品。论文最后分析了AlphaReadabilityChinese在数字人文、国际中文教育、新闻传播、信息科学、经济/金融等学科领域的应用前景。

关键词:汉语;文本可读性;金庸;古龙;数字人文

中图分类号:H0　文献标识码:A　文章编号:1004 - 6038 (2024) 01 - 0083 - 11

1. 引言

文本可读性指的是文本的难度或易读性(Crossley et al. 2019; Flesch 1948)。数十年来,文本可读性一直受到研究者关注,其中一个重要原因在于其是文本的重要语言特征,且具有多样的应用场景,因此被广泛用于信息科学(Lei & Yan 2016)、新闻传播学(Graefe et al. 2018)、经济学/金融学(Aldoseri & Melegy 2023)等多学科领域研究。

研究者很早就尝试开发文本可读性的指标和工具,用于自动测量文本的可读性。早期经典的文本可读性指标多基于文本的简单语言特征,其中常见特征包括文本的句子长度、单词长度(单词的字母数或音节数)、难词(低频词)等。比如,Flesch (1948)开发的易读性指标(the Flesch Reading Ease)可能是目前使用最广泛的文本易读性指标,涉及文本的句子总数、单词总数、音节总数等特征(见公式①)。又如,McLaughlin (1969)开发的简单可读性指标(the Simple Measure of Gobbledygook, SMOG)只包括多音节单词数一个特征(多音节单词指

* **基金项目:**本文为上海外国语大学重大攻关项目"面向国际中文教育的语料库建设、系统开发及创新研究"(项目编号:23ZD011)的阶段性研究成果。

作者简介:雷蕾,教授,博士,研究方向:语料库语言学,语言数字人文;韦瑶瑜,副教授,博士,研究方向:二语习得,语料库语言学;刘康龙,副教授,研究方向:语料库语言学,翻译研究。第一作者邮箱:leileicn@126.com

三个或以上音节单词，见公式②。再如，Chall和Dale (1995)开发的新可读性公式(the New Dale-Chall Readability Formula)包括平均句长、单词总数、难词总数等文本特征，他们将难词定义为四年级3,000词表以外的单词，见公式③。

①易读性指标(Flesch 1948)

$$206.835-1.015\times\frac{\text{单词总数}}{\text{句子总数}}-84.6\times\frac{\text{音节总数}}{\text{单词总数}}$$

②简单易读性指标 (McLaughlin 1969)

$$3+\sqrt{\text{多音节单词数}}$$

③新可读性公式 (Chall & Dale 1995)

$$64-0.95\times100\times\frac{\text{难词总数}}{\text{单词总数}}-0.69\times(\text{平均句长})$$

上述经典可读性指标计算简单，因此深受研究者欢迎，历经数十年仍被广泛运用于各领域研究。然而，由于这些指标只包括文本的音节、单词、句子数等语言形式特征，而没有考虑句法、语义等其他影响文本可读性的重要因素，也受到诸多质疑。因此，很多学者试图完善文本可读性计算指标体系，基于更成熟的技术手段，开发新的可读性指标和工具。这些新工具不仅包括词汇层面的语言特征，也包括句法和语义等层面的特征。比如，Crossley等(2019)综合考虑文本的词汇、句法、连贯、认知等因素，探索文本可读性指标。研究发现，文本实词的形象性(imageability)、字符熵、词汇的具体性(concreteness)和语义独特性等因素是影响阅读理解和阅读速度的重要因素。基于研究发现，Crossley等(2019)建议，可读性指标和工具开发可考虑加入上述词汇、句法、连贯、认知等层面的因素。由于Crossley等(2019)开发的算法过于复杂，Crossley等(2023)进一步改进Crossley等(2019)算法，采用句嵌套和BERT模型等来测量文本可读性。随着汉语语言习得研究的发展，特别是针对汉语一语或二语学习者文本或阅读材料可读性研究的深入，汉语文本可读性指标的开发也受到研究者的关注。与英语文本可读性指标开发类似，早期的汉语文本可读性指标开发相关研究，大多也只考虑字、词、句等层面长度或与频次相关的语言特征。比如，Yang (1971)是较早探索汉语文本可读性的研究。该研究基于汉字、单词、句子等层面共39个特征，通过多元回归分析，构建了可读性公式，其公式最终包括难词比、完整句子数、汉字平均笔画数等三个特征。又如，左虹和朱勇(2014)、王蕾(2017)等也通过计算汉语文本的汉字平均笔画数、难词比率、简单词比率、虚词数、平均句长、文本长度等表层语言特征来开发可读性公式。

近年来，汉语文本可读性指标开发相关研究也逐渐加入其他语言特征。比如，朱君辉等(2022)等参考《国际中文教育中文水平等级标准》中的汉字、词汇和语法点级别等信息，提取572条语法特征，并依据语法特征对国际中文教育教材不同级别文本难度进行自动分级，发现语法特征可以有效区分不同难度级别的文本。吴思远等(2020)采用汉字、词汇、句法和篇章四个维度共104项语言特征指标，对中小学12个年级语文课本文本进行自动分级实验，发现词汇维度特征对于文本难度的预测准确率最高，所有语言特征指标中预测力最强的是汉字熟悉度、汉字多样性、词汇多样性、短语句法结构复杂度、词汇熟悉度等指标。程勇等(2020)也对中小学12个年级语文课本文本难度自动分级进行研究，发现字频、词义丰富度、

连词比例等特征是区别文本难度级别的关键因素。

文本可读性指标不仅可以用于上述语言教学与习得研究,在其他学科领域也得到了广泛应用。比如,经济/金融领域,上市公司年报的可读性与企业其他经济指标关系的研究发现企业年报的可读性是企业信息披露质量的重要指标,其受到企业数字化转型的影响,企业数字化转型程度越高,企业年报可读性越高(王海芳等 2022)。又如,图书情报科学领域研究者关注学术文本可读性与其影响力的关系,发现学术论文文本可读性越低,其被其他论文引用的次数越高,其在学界的影响力越大(刘宇等 2023);有意思的是,与之相反,学术论文文本可读性越低,其在社交媒体的传播率越低,其受公众关注的程度也越低(欧桂燕等 2023)。在近年兴起的数字人文领域(胡开宝、王晓莉 2022; 雷蕾 2023),研究者采用文本可读性指标中重要的语言特征(如词长、句长、词汇丰富度等)分析文学作品风格,识别文学作品的作者。比如,刘颖和肖天久(2014)对比分析了金庸和古龙小说作品的段落长度、句长、虚词频率等特征,研究发现,金庸小说文本可读性较古龙小说文本要弱。该研究还依据文本中虚词和高频词频率对两位作家的作品做聚类分析,发现虚词频率和高频词频率可以较好地聚类两位作家的作品。

现有的文本可读性指标开发相关研究小结如下。现有的可读性研究大多数针对英语文本可读性指标和工具的开发,针对汉语文本的可读性研究,虽已有数十年历史,但相较于英语文本相关研究,汉语文本的可读性研究尚处于起步阶段,其主要体现在如下几个方面。首先,从汉语文本可读性指标的维度来看,现有研究大多采用词汇和句法等表层维度的简单指标来计算汉语文本的可读性(左虹、朱勇 2014)。近年来,少数研究开始将篇章和语义等维度纳入汉语文本可读性指标(程勇等 2020)。其次,从汉语文本可读性指标的计算方法来看,现有研究大多采用长度、频次、比率等简单计算方法(王蕾 2017),而极少采用更成熟、稳健的算法来计算文本可读性指标,如信息熵(Liu et al. 2022)。最后,现有汉语文本可读性研究,大多面向国际中文教育教材和学习者文本开发指标并进行相关研究 (吴思远等 2020; 朱君辉等 2022),这些指标在其他领域研究的适用性尚有待验证。

本研究旨在报告我们开发的汉语文本可读性工具,我们将之命名为 AlphaReadabilityChinese (简称 ARC)[①]。ARC 拟解决上述现有汉语文本可读性指标和工具的几个局限性:一是 ARC 拟纳入词汇、句法、语义等多个维度的指标;二是 ARC 拟采用更成熟、稳健的计算方法来开发可读性指标;三是 ARC 是通用的汉语可读性指标和工具,其不应仅局限于国际中文教育等教学和习得的应用场景,而应广泛用于多学科、多领域场景的研究。本文第 2 小节将着重介绍 ARC 的主要指标构成,在第 3 小节,我们将以数字人文研究/文学研究为例,通过具体实验测定 ARC 的有效性。

2. 汉语文本可读性工具开发

本研究开发的汉语文本可读性工具 ARC 包括词汇、句法、语义三个维度,共九个指标,如表 1 所示。下文我们将详细描述 ARC 的各指标。

表1 汉语文本可读性工具维度及指标

维度	指标名称	指标代码	计算方法	备注
词汇	词汇丰富度	lexical_richness	单词的熵值	正比*
句法	句法丰富度	syntactic_richness	依存关系的熵值	正比
语义	名词语义精确度	semantic_accuracy_n	名词义项数的平均数	反比**
语义	动词语义精确度	semantic_accuracy_v	动词义项数的平均数	反比
语义	名词与动词语义精确度	semantic_accuracy_n_v	名词与动词义项数的平均数	反比
语义	实词语义精确度	semantic_accuracy_c	实词义项数的平均数	反比
语义	语义丰富度	semantic_richness_n	名词概率之和	正比
语义	语义清晰度	semantic_clarity_n	名词概率的偏度	反比
语义	语义噪音	semantic_noise_n	名词概率的峰度	正比

注:* 指标值越大,文本难度越大,文本越难读;** 指标值越大,文本难度越小,文本越易读。

(1) 词汇丰富度

我们按照公式④通过计算文本所有单词的熵值来计算文本的词汇丰富度。熵是信息科学的概念,该统计量测量的是特定系统信息的随机性或不确定性(Shannon 1948)。熵值也被广泛用于测量文本信息的不确定性,如测量文本词汇的不确定性或词汇丰富度(Liu et al. 2022)。文本词汇丰富度的值越大,说明文本所使用的单词越不确定,文本的词汇越富于变化,则文本阅读难度越大。

④词汇丰富度计算公式

$$\text{Entropy} = -\sum_{i=1}^{n} P_i log_2 P_i$$

其中,P_i为单词在文本中出现的概率或相对频次,n为文本包含的所有单词总数[②]。

(2) 句法丰富度

与计算词汇复杂度类似,我们通过计算文本所有依存关系的熵值来计算文本的词汇丰富度(同公式④)。句子的依存关系代表的是句子中词与词之间的句法关系,也是文本或句子句法结构及句法关系的重要体现(Lei & Wen 2020)。文本句法丰富度的值越大,说明文本的依存关系或句法结构越不确定,文本的句法越富于变化,则文本越难读。

(3) 语义精确度

文本中的语义主要通过名词、动词、形容词、副词等实词来实现,我们通过计算文本中名词、动词、名词与动词、所有实词义项数的平均数来计算文本的语义精确度(见公式⑤)。以计算名词的语义精确度为例,其计算方法为名词义项数总和与名词总数之商。值得注意的是,越高频越简单的实词,其义项数越大,语义越不精确;与之相反,越低频越难的实词,其义项数越小,语义越精确(McNamara et al. 2015)。因此,文本语义精确度的值越大,其用词越简单,则文本越易读;反之,语义精确度值越小,其用词越难,则文本越难读。

⑤语义精确度计算公式

$$SA = \frac{\sum_{i=1}^{n} S_i}{n}$$

以名词为例，S_i为文本中名词的义项数，n为文本的名词总数。

(4) 语义丰富度

文本的话题丰富程度也可能影响文本的可读性。文本的话题越丰富，其语义越丰富，则文本的可读性可能越低。Lee等(2021)先对文本进行LDA话题建模以提取文本的话题，然后通过计算话题概率分布与话题排序之积的总和来计算文本语义丰富度。该算法可能存在三个问题。一是，虽然LDA话题建模算法使用广泛，但由于其提取的话题数及话题展示方式等方面的问题而受到学界诟病(Lei et al. 2020)。二是，LDA话题建模算法更适合长文本话题的建模或提取，如果文本较短(如学习者写作文本)，则建模效果会受到影响。三是，以话题概率排序来调整文本丰富度的权重，似无理论或实践依据。为解决上述问题，我们采用简化的语义丰富度算法。由于文本的话题大多由名词或名词词组构成(Lei et al. 2020)，我们通过计算文本中名词出现概率之和来计算文本的语义丰富度(见公式⑥)。语义丰富度值越大，说明文本的话题越丰富，文本的可读性越低。

⑥语义丰富度计算公式

$$SR = -\sum_{i=1}^{n} P_i$$

其中，P_i为名词在文本中出现的概率或相对频次，n为文本的名词总数。

(5) 语义清晰度

Lee等(2021)采用LDA话题建模提取的文本话题分布概率的偏度(skewness)来计算文本的语义清晰度。偏度测量的是数据观测分布与数据正态分布的偏离方向和程度。Lee等(2021)认为，文本中话题分布概率越向右偏离正态分布，则其话题越集中，语义越清晰。我们采用Lee等(2021)开发的偏度计算方法(见公式⑦；该公式与标准偏度算法略有不同)，计算文本中名词出现概率的偏度，也就是文本的语义清晰度。语义清晰度值越大，说明文本以名词为代表的话题越集中，其语义越清晰。

⑦语义清晰度计算公式

$$SC = \frac{1}{N}\sum_{i=1}^{n} (max(P) - P_i)$$

其中，P_i为名词在文本中出现的概率或相对频次，N为文本的单词总数，n为文本的名词总数。

(6) 语义噪音

Lee等(2021)采用LDA话题建模提取的文本话题分布概率的峰度(kurtosis)来计算文本的语义噪音。峰度测量的是随机变量概率分布的陡峭程度以及概率分布的尾度(tailness)。Lee等(2021)认为，文本中话题分布概率的峰度越大，其尾度越厚，则话题分布越偏向于不重要的话题，文本的语义噪音越大。我们采用与Lee等(2021)相同的峰度公式(见公式⑧)，计算文本中名词出现概率的峰度/尾度或语义噪音。语义噪音值越大，说明文本以名词为代表的话题越偏向不重要的话题，其语义噪音越大。

⑧语义噪音计算公式

$$SN = n \times \frac{\sum_{i=1}^{n} (P_i - \bar{P})^4}{(\sum_{i=1}^{n} (P_i - \bar{P})^2)^2}$$

其中，P_i为名词在文本中出现的概率或相对频次，n为文本的名词总数。

从工具的技术实现来看，在开发汉语文本可读性工具ARC时，我们采用哈尔滨工业大学开发的自然语言处理工具LTP对汉语文本进行分词处理和依存句法分析(Che et al. 2021)。另外，我们采用清华大学开发的概念和语义知识库OpenHowNet来计算汉语文本中名词等实词的义项数。

3. 验证汉语文本可读性工具：文学作品研究案例

为验证汉语文本可读性工具ARC是否可以准确测量文本可读性，我们将之用于分析金庸和古龙小说文本的可读性。本小节我们汇报该研究案例的结果。

3.1 研究目的、数据与方法

金庸和古龙是当代著名的、最受欢迎的武侠小说大家，他们的武侠小说作品在海内外传播广泛，具有重要影响力(严家炎 2019)。金庸小说更是由于其思想意识品质和传播接受度，“登堂入室”进入《百年中国文学经典》(钱理群、谢冕 1996)，从而引起了“文学史重写”与“文学经典”之争(陈夫龙 2017)。除了其思想和内容方面存在“大侠”与“游侠”等不同以外(韩云波 2017)，金庸和古龙的小说作品在语言特征上也存在诸多差异。目前，对金庸和古龙小说作品做详细对比分析的仅有刘颖和肖天久(2014)一项研究。上述研究从段落长度、句子长度、高频词、虚词、标点符号、文本从众性、句子破碎度等语言特征视角对比分析了金庸和古龙的作品，结果发现，古龙小说的用词更富于变化，但其较金庸小说更易读。另外，该研究还通过高频词、词类等语言特征对两位作家作品做聚类分析，发现上述语言指标可以很好地区分二者作品，说明二者作品语言风格存在较大差异。

本研究案例运用汉语文本可读性工具ARC对金庸和古龙小说文本进行可读性分析，比较二者之间可能存在的差异，以验证汉语文本可读性工具ARC的效度。另外，本研究案例采用与刘颖和肖天久(2014)相同的数据样本，以方便将本研究结果与之前的研究结果进行对比。具体来说，我们选取了金庸的《射雕英雄传》《神雕侠侣》《倚天屠龙记》《天龙八部》《笑傲江湖》《鹿鼎记》等六部小说，以及古龙的《大旗英雄传》《武林外史》《绝代双骄》《楚留香传奇》《小李飞刀》《陆小凤传奇》等六部小说。

本研究案例回答如下问题：

(1) 金庸和古龙小说作品在可读性特征上是否存在差异？

(2) 可读性指标是否可以区分金庸和古龙小说作品？

我们采用如下研究方法处理数据，以回答上述研究问题。首先，我们运用汉语文本可读性工具ARC分别对金庸和古龙12部小说进行可读性分析。其次，我们采用T检验对两位作家作品的可读性指标做对比分析，以回答研究问题1。最后，我们基于可读性指标将两位作家的12部作品做聚类分析，以回答研究问题2。关于聚类分析，我们先对ARC所包含的九个可读性指标值作归一化处理，并基于归一化后的可读性指标值来计算文本的欧几里得距离。然后，计算聚类系数(agglomerative coefficient)，以确定聚类连结方法(clustering linkage method)。由于Ward方法的聚类连结系数最大，我们确定以Ward方法连结聚类。最后，通过层次聚类算法对12部小说文本进行聚类。

3.2研究结果与讨论

本小节我们汇报研究案例的结果,并对结果进行简要讨论。金庸和古龙作品文本可读性指标描述性统计和T检验结果见表2和表3。

首先,从词汇丰富度指标来看,金庸作品的词汇丰富度指标的平均值大于古龙作品的平均值指标,T检验结果显示两位作家作品在该指标上具有显著差异。也就是说,金庸作品的词汇更富于变化,其文本可读性弱于古龙作品文本的可读性。有意思的是,我们关于词汇丰富度的发现与刘颖和肖天久(2014)的结果相左。刘颖和肖天久(2014)通过词长变化程度来对比分析两位作家作品的文本,发现古龙作品的词长变化程度更大,因此词汇使用更丰富。刘颖和肖天久(2014)认为,古龙中后期作品语言风格变化较大,较短篇幅作品文本(如《小李飞刀》)更具诗化特点,因此词长变化较大。然而,基于词长变化程度来测量文本词汇丰富度可能存在一定局限性。一是汉语文本书写时没有自然的单词边界,需要借助自然语言处理相关工具来切分单词,因此单词长度取决于单词切分的颗粒度,如果颗粒度较粗,则词长较长,反之亦然。二是汉语单词多由一或两个汉字构成(据Chen和Liu (2016)统计,由一个或两个汉字构成的汉语单词占比超过80%),少数由三个或四个汉字构成(Chen & Liu 2016),因此,汉语文本的词长变化不大,很难成为具有显著意义的语言特征指标。三是,尽管很多研究将词长作为作者识别和写作风格研究的指标(Lian & Li 2021),但也有学者认为,由于词长并不是作者写作风格的显著特征,因此单纯依靠词长信息以识别作者,其准确率不高(Zheng & Jin 2023)。综上,词长变化程度仅能体现文本词长的变化,而汉语文本大多以一或两个汉字构成,因此词长似乎并不能反映汉语文本的词汇丰富度。本研究采用单词熵值算法,其体现的是文本词汇使用的不稳定性或变化程度,似能更好地反映文本词汇丰富度。本研究结果发现,金庸作品的词汇丰富度显著大于古龙文本的词汇丰富度,故其作品的可读性弱于古龙作品。这一发现也证明了刘颖和肖天久(2014)古龙作品可读性更高的观点。

与词汇丰富度类似,金庸作品的句法丰富度指标的平均数也大于古龙作品的平均数指标,T检验结果表明两个指标均具有显著差异,说明金庸作品的句法结构更富于变化,因此,从句法丰富度视角来看,金庸作品的可读性也弱于古龙作品。

表2 金庸和古龙作品文本可读性指标描述性统计

作者	指标	最小值	最大值	中位数	平均值	标准差
金庸	词汇丰富度	6.69	6.90	6.82	6.79	0.08
金庸	句法丰富度	2.17	2.18	2.18	2.18	0.01
金庸	名词语义精确度	3.97	4.2	4.01	4.03	0.09
金庸	动词语义精确度	9.53	9.91	9.65	9.67	0.15
金庸	名词与动词语义精确度	7.41	7.73	7.5	7.55	0.14
金庸	实词语义精确度	7.34	7.71	7.43	7.48	0.14
金庸	语义丰富度	0.13	0.14	0.13	0.14	0.00
金庸	语义清晰度	0.04	0.05	0.05	0.05	0.00
金庸	语义噪音	3,431.34	4,907.38	4,114.71	4,137.47	742.22
古龙	词汇丰富度	6.14	6.54	6.29	6.29	0.14

（续表）

作者	指标	最小值	最大值	中位数	平均值	标准差
古龙	句法丰富度	2.15	2.17	2.17	2.16	0.01
古龙	名词语义精确度	4.21	5.34	5.03	4.92	0.45
古龙	动词语义精确度	9.82	10.46	10.2	10.19	0.24
古龙	名词与动词语义精确度	7.75	8.69	8.42	8.36	0.35
古龙	实词语义精确度	7.46	8.16	8.02	7.92	0.26
古龙	语义丰富度	0.11	0.13	0.11	0.11	0.01
古龙	语义清晰度	0.06	0.09	0.09	0.08	0.02
古龙	语义噪音	2,567.6	5,958.87	3,408.65	3,804.96	1,400.65

其次，从语义精确度系列指标来看，金庸作品文本的名词、动词、名词与动词、实词的语义精确度值均显著低于古龙作品。也就是说，金庸作品更多使用更难的词汇，其使用的词汇从语义来看更加精确，因此其文本更难读，文本可读性弱于古龙作品。

最后，从语义丰富度指标来看，金庸作品话题概率分布更大，其作品比古龙作品话题更丰富，也使得其文本可读性较弱。从语义清晰度指标来看，也有类似发现，即古龙作品文本语义清晰度值大于金庸作品，说明其话题分布概率更向右偏离正态分布，其话题更集中，语义更清晰。

汉语文本可读性工具ARC测量的所有指标中，只有语义噪音指标在两位作家的作品文本中没有显著差异。语义噪音指标测量的是文本话题偏向于不重要话题的程度或偏离重要话题的程度。金庸和古龙都是武侠小说大家，虽然他们的作品内容、篇幅、语言风格等具有较大差异，但两位作家的作品均聚焦于小说主线展开，因此他们作品文本的语义噪音没有显著差异。

表3 金庸和古龙作品文本可读性指标T检验结果

指标	T值	自由度	P值
词汇丰富度	7.80	8.08	0.00
句法丰富度	2.71	7.84	0.03
名词语义精确度	−4.72	5.38	0.00
动词语义精确度	−4.47	8.60	0.00
名词与动词语义精确度	−5.35	6.50	0.00
实词语义精确度	−3.65	7.81	0.01
语义丰富度	6.86	6.14	0.00
语义清晰度	−5.24	5.59	0.00
语义噪音	0.51	7.60	0.62

金庸和古龙12部小说的层次聚类结果如图1所示。研究结果表明，通过汉语文本可读性工具ARC的九个指标可以很好地将两位作家的作品聚类成两个大类。一方面，这一结果与刘颖和肖天久(2014)类似，他们基于虚词、高频词、N元语法等语言特征也将两位作家的作品聚成两大类，说明金庸和古龙作品的语言特征的确存在显著差异。另一方面，这一结果也

从另一视角证明了本研究开发的汉语文本可读性工具ARC及其所包含指标可用于区分不同作者文本的可读性,证明了汉语文本可读性工具ARC具有较为理想的效度,也证明了其可用于语言数字人文或作者识别研究领域的适用性。

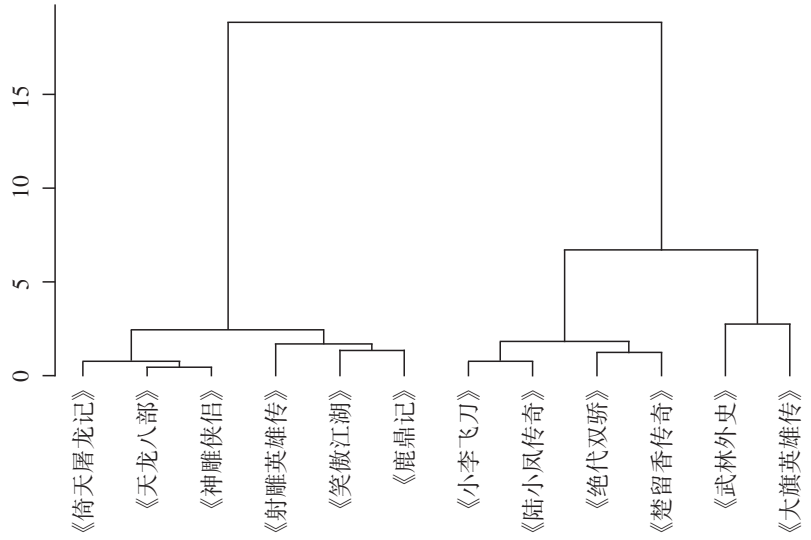

图1 金庸和古龙作品的聚类分析

4. 启示与展望

本文报告了我们开发的汉语文本可读性工具AlphaReadabilityChinese及其研究案例。该工具可从词汇、句法、语义等多个维度测量汉语文本的可读性。我们将该工具用于金庸和古龙作品的可读性分析,发现两位作家的作品在绝大多数可读性指标上具有较大差异,金庸作品的可读性显著弱于古龙作品。另外,该工具所包含的可读性指标可以很好区分两位作家的作品。

与现有的汉语文本可读性工具相比,ARC具有如下显著特征。首先,ARC包括词汇、句法、语义等多个维度的指标,构造了较全面的多维度指标体系。除了现有相关研究关注的词汇和句法维度的指标以外(朱君辉等 2022),ARC特别关注语义维度的文本特征,包括了语义精确度、语义丰富度、语义清晰度、语义噪音等多个语义维度指标。语义维度特征对于文本可读性测量的准确与否至关重要,基于文本语义相似度的可读性测量准确率也显著高于基于传统的词汇和句法维度的准确率(Crossley et al. 2023)。因此,本研究开发的汉语文本可读性工具特别关注语义维度的可读性指标,部分弥补了现有研究对语义维度可读性关注不足的缺憾。

其次,ARC采用更成熟、稳健的计算方法来开发可读性指标。从我们掌握的可读性相关文献来看,本研究提出的基于句法依存熵值的句法丰富度算法、基于名词和动词等实词平均义项数的语义精确度算法、基于名词话题概率分布的语义丰富度/清晰度/噪音算法等,均为类似指标首次用于汉语文本可读性测量的案例,是汉语文本可读性指标和工具开发的有益尝试。

再次,从应用场景来看,我们开发的汉语文本可读性工具ARC为通用的汉语可读性指标和工具,其可用于多学科、多领域的研究。本研究将之用于语言数字人文/文学作品语言

特征的分析，从研究案例的结果来看，ARC可以有效区分金庸和古龙作品文本的可读性，该结果也为ARC测量汉语文本可读性的有效性提供了证据。今后研究还可尝试将ARC用于其他学科的相关研究。比如，可将之用于语言数字人文领域的文学或历史作品文本特征分析或作者识别等研究(雷蕾 2023)、国际中文教育等教学材料难度的自动分级、以汉语为一语或二语课内及课外学习材料开发、学习者写作文本自动评分等相关研究(吴思远等 2020; 朱君辉等 2022)、新闻/传播领域的文本特征与其传播效果或传播接受度关系等研究(Graefe et al. 2018)、信息学科领域的科学论文文本特征与其影响力关系等研究(Lei & Yan 2016)，或经济学/金融学领域文本特征与经济/金融相关指标关系的研究等(Aldoseri & Melegy 2023)。

最后，现有可读性分析工具往往包括几十甚至百余个可读性指标(朱君辉等 2022)。本研究在开发ARC时，采用了与现有可读性工具不同的思路，即我们既要求ARC囊括词汇、句法、语义等多个维度的重要指标，又要求ARC指标数量保持在一定数量。上述少而精的思路试图在指标数量、指标质量及工具运行效率上取得平衡，即在确保ARC指标体系全面、有效的同时，又兼顾工具运行和数据处理的高效性。

本研究也存在一定局限性。首先，本研究没有包括语篇连贯等维度的指标，今后研究可从连贯等维度入手进一步完善ARC指标体系。另外，今后研究可尝试运用ARC指标开发汉语文本可读性回归方程，或结合新兴算法或技术，开发通用或专门用途/垂直领域的汉语文本可读性大语言模型(Crossley et al. 2023)，以进一步开拓汉语文本可读性研究边界。最后，我们仅基于文学文本验证了ARC的效度，今后研究可在其他领域(如语言数字人文、国际中文教育、新闻传播、信息科学、经济学/金融学等)进一步验证ARC的适用性。

注释：

①我们开发了用于英语文本可读性研究的指标和工具，并将之命名为AlphaReadability。因此，依据我们开发系列工具的命名惯例，将汉语文本可读性指标和工具命名为AlphaReadabilityChinese。

② 本文所述文本的单词数，均为文本形符(token)数。

参考文献：

[1] Aldoseri, M. & M. Melegy. 2023. Readability of annual financial reports, information efficiency, and stock liquidity: Practical guides from the Saudi business environment[J]. *Information Sciences Letters*, (12): 813-821.

[2] Chall, J. & E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*[M]. Boston: Brookline Books.

[3] Che, W., Y. Feng, L. Qin & T. Liu. 2021. N-LTP: An open-source neural language technology platform for Chinese[A]. In H. Adel & S. Shi (eds.). *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*[C]. Punta Cana: Association for Computational Linguistics.

[4] Chen, H., & H. Liu. 2016. How to measure word length in spoken and written Chinese[J]. *Journal of Quantitative Linguistics*, (23): 5-29.

[5] Crossley, S., S. Skalicky & M. Dascalu. 2019. Moving beyond classic readability formulas: New methods and new models[J]. *Journal of Research in Reading*, (42): 541-561.

[6] Crossley, S., J. Choi, Y. Scherber & M. Lucka. 2023. Using large language models to develop readability formulas for educational settings[J]. *Communications in Computer and Information Science*, (1831): 422-427.

[7] Flesch, R. 1948. A new readability yardstick[J]. *Journal of Applied Psychology*, (32): 221-233.
[8] Graefe, A., M. Haim, B., Haarmann & H. Brosius. 2018. Readers' perception of computer-generated news: Credibility, expertise, and readability[J]. *Journalism*, (19): 595-610.
[9] Lee, B., Y. Jang & J. Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features[A]. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*[C]. Punta Cana: Association for Computational Linguistics.
[10] Lei, L., Y. Deng & D. Liu. 2020. Examining research topics with a dependency-based noun phrase extraction method: A case in accounting[J]. *Library Hi Tech*, (41): 570-582.
[11] Lei, L. & J. Wen. 2020. Is dependency distance experiencing a process of minimization?: A diachronic study based on the State of the Union addresses[J]. *Lingua*, (239): 1-13.
[12] Lei, L. & S. Yan. 2016. Readability and citations in information science: Evidence from abstracts and articles of four journals (2003-2012)[J]. *Scientometrics*, (108): 1155-1169.
[13] Lian, F. & Y. Li. 2021. Word length distribution in German texts during the 17th–19th century[J]. *Journal of Quantitative Linguistics*, (28): 117-137.
[14] Liu, K., Z. Liu & L. Lei. 2022. Simplification in translated Chinese: An entropy-based approach[J]. *Lingua*, (275):1-14.
[15] McLaughlin, G. 1969. SMOG grading: A new readability formula[J]. *Journal of Reading*, (12): 639-646.
[16] McNamara, D., S. Crossley, R. Roscoe, L. Allen & J. Dai. 2015. A hierarchical classification approach to automated essay scoring[J]. *Assessing Writing*, (23): 35-59.
[17] Shannon, C. 1948. A mathematical theory of communication[J]. *Bell System Technical Journal*, (27): 379-423.
[18] Yang, S. 1971. *A Readability Formula for Chinese Language*[D]. Madison: The University of Wisconsin.
[19] Zheng, W. & M. Jin. 2023. Is word length inaccurate for authorship attribution?[J]. *Digital Scholarship in the Humanities*, (38): 875-890.
[20] 陈夫龙 . 2017. 金庸小说经典化之争及其反思[J]. 小说评论, (5): 42-49.
[21] 程勇 徐德宽 董军 . 2020. 基于语文教材语料库的文本阅读难度分级关键因素分析与易读性公式研究[J]. 语言文字应用, (1): 132-143.
[22] 韩云波 . 2017. 从"前金庸"看金庸小说的历史地位[J]. 浙江学刊, (2): 76-87.
[23] 胡开宝 王晓莉 . 2022. 数字人文视域下翻译研究:现状、问题与前景[J]. 外语与外语教学, (6): 111-121.
[24] 雷蕾 . 2023. 语言数字人文:"小帐篷"理论框架[J]. 外语与外语教学, (3): 63-73.
[25] 刘颖 肖天久 . 2014. 金庸与古龙小说计量风格学研究[J]. 清华大学学报(哲学社会科学版), (5): 135-147+179.
[26] 刘宇 伍丹炜 叶继元 . 2023. 文本可读性与学术论文的影响力:基于图书情报学的实证研究[J]. 中国图书馆学报, (5): 111-127.
[27] 欧桂燕 陈佩 吴江 . 2023. 学术文本可读性特征对 Altmetrics 的影响研究——以 Web of Science 论文摘要数据为例[J]. 图书情报工作, (4): 102-113.
[28] 钱理群 谢冕 . 1996. 百年中国文学经典[M]. 北京:北京大学出版社 .
[29] 王海芳 姜道平 许莹 . 2022. 数字化转型能否提高信息披露质量?——基于年报可读性的研究[J]. 管理现代化, (2): 58-65.
[30] 王蕾 . 2017. 初中级日韩学习者汉语文本可读性公式研究[J]. 语言教学与研究, (5): 15-25.
[31] 吴思远 于东 江新 . 2020. 汉语文本可读性特征体系构建和效度验证[J]. 世界汉语教学, (1): 81-97.
[32] 严家炎 . 2019. 金庸小说成就之我见[J]. 浙江学刊, (6): 14-20.
[33] 朱君辉 刘鑫 杨麟儿 王鸿滨 杨尔弘 . 2022. 汉语语法点特征及其在二语文本难度自动分级研究中的应用[J]. 语言文字应用, (3): 87-99.
[34] 左虹 朱勇 . 2014. 中级欧美留学生汉语文本可读性公式研究[J]. 世界汉语教学, (2): 263-276.