

学术文本可读性和复杂度评价研究*

陈练文¹ 李信² 赵超烨²

(1. 武汉大学文学院中国语情与社会发展研究中心, 武汉 430072; 2. 武汉大学信息管理学院, 武汉 430072)

摘要: 本文以计算机和图书情报学科的学术文本作为研究对象, 综合SMOG、FKG、ARI、FOG 4种可读性评价公式和句法复杂度分析工具L2SCA, 从学科、学术文本结构位置 and 不同影响因子期刊等方面出发, 对学术文本的可读性和复杂度进行比较分析。研究结果表明, 不同学科学术文本之间的可读性差异不明显, 而学术文本正文部分的可读性都显著低于摘要部分; 此外, 在复杂度方面, 正文在大多数指标上均比摘要的复杂度高。

关键词: 学术文本; 文本可读性; 句法复杂度; 可读性公式; 复杂度测评

中图分类号: TP391.1; G35

DOI: 10.3772/j.issn.1673-2286.2018.05.009

1 学术文本可读性研究现状

学术论文作为科研活动的主要成果展现形式, 是科研人员学术水平和研究工作最直接的体现, 同时也是反映高等学校和科研院所科研水平、办学能力的重要指标^[1]。学术文本的复杂度直接关系到学术成果的传播、展示和科研人员间的沟通交流, 是学术论文评价不可忽略的部分。然而, 调研发现, 目前国内外关于学术文本复杂度评估的研究很少, 基本处于空白状态^[2]。因此, 本文拟从文本可读性和句法复杂度两个角度对学术文本的复杂度进行定量研究, 以期完善学术论文评价体系提供借鉴。

可读性, 也被称为易读度或易读性, 指文本易于阅读和理解的程度或性质^[3], 主要用来评价阅读材料能被读者阅读和理解的容易程度^[4]。影响可读性的因素主要有句子的平均长度、陌生词汇的数量、所用语法的复杂程度^[5]。虽然可读性研究已有一定的历史, 但相关研究主要关注的是可读性与语言教育的关系, 如Ortega^[6]和Persson^[7]对语言教育材料可读性的相关研究。有关学术文本可读性的研究还较少, 主要集中在可读性与其学术影响力的关系上, 如Armstrong^[8]、Stremersch等^[9]对Dr. Fox phenomenon现象的证明, Sawyer等^[10]发现获奖的学术论文更容易阅读, Dolnicar等^[11]对旅游期刊论文进

行可读性研究, Sun^[12]认为低可读性文本的剽窃概率更高等。学术文本不同结构通常有不同的语义功能^[13], 因此, 一部分学者着力评估学术文本某一方面的可读性, 如Plavén-Sigray等^[14]指出学术文本可读性随时间逐渐下降, Gazni^[15]和Lei等^[16]认为学术文本摘要比正文的可读性低。句法复杂度指语言产出中句法结构的多样性及复杂性^[17]。句法复杂度关系到语法结构的掌握和产出, 虽然已有学者意识到学术论文评价在内容认知层面的不足^[18-19], 但将可读性、复杂度和学术文本相结合的学术论文评价实践研究还较少^[20]。

通过以上分析可知, 学术文本的可读性和句法复杂度研究尚未获得足够重视, 相关成果较少。相对来说, 可读性研究比句法复杂度研究成果更多, 但都存在以下不足: ①现有研究仅针对单一学科的学术文本, 缺乏不同学科、不同影响力学术文本的对比研究; ②以往研究大多采用单一的公式或指标, 得到的结果可靠性不高; ③已有研究成果的评估粒度主要停留在文章层面, 缺乏对文章不同结构(如摘要、正文等)、组成成分(如句子)的细粒度评估; ④现有研究主要针对特定人群(如外语学习者、某国学者)的学术文本, 研究成果缺乏普遍性。因此, 本文拟从文本可读性和句法复杂度两个方面, 采用多种评价公式和工具, 全方位、多角度对学术文本的可读性和句法复杂度进行分析。

*本研究得到国家社会科学基金重大项目“基于认知计算的学术论文评价理论与方法研究”(编号: 17ZDA292)资助。

2 数据来源和研究方法

2.1 数据来源

本文以计算机和图书情报学科的学术论文为例,探讨不同类型文本可读性的异同。2016年7月1日—8月31日,自主采集Springer数据库中113本计算机(Computer Science, CS)学科和6本图书情报(Library and Information Science, LIS)学科期刊论文全文数据,共得到全文HTML网页数据294 332条,其中图书情报学科数据11 956条。

针对每篇学术论文,区分论文摘要(ABS)和正文(PASS),最终得到4组数据,CS_ABS、CS_PASS、LIS_ABS和LIS_PASS分别表示计算机学科学术论文摘要、计算机学科学术论文正文、图书情报学科学术论文摘要及图书情报学科学术论文正文。以上数据的统计信息见表1。

表1 学术文本数据的统计信息

数据名称	统计指标	
	文本数量/条	平均文本长度(以句子数计算)
CS_ABS	236 237	6.346 4
CS_PASS	206 781	159.088 2
LIS_ABS	5 795	6.039 2
LIS_PASS	6 180	161.504 4

2.2 研究方法

2.2.1 可读性评估方法与指标

本文从文本类型、所属学科、结构位置和期刊等级的视角出发,通过对学术文本的可读性进行定量测评、比较分析和相关性分析,以探索学术论文可读性规律。主要使用Simple Measure of Gobbledygook (SMOG)、Flesch-Kincaid Grade Level (FKG)、Automated Readability Index (ARI) 和Gunning-FOG (FOG) 这4个可读性测评公式进行评估。

另外,本文将每一个数据集的文本可读性定义为数据集中所有文本可读性得分的平均值。需要注意的是,SMOG公式只适用于文本长度大于30的文本,因此除LIS_PASS和CS_PASS外,其余文本均不计算SMOG得分。

2.2.2 句法复杂度评估方法与指标

本文对句法复杂度的测量工具是L2SCA^[17],主要从平均句长、平均小句长、平均T单位长等14个指标对文本的句法复杂度进行评估。实验数据与文本可读性分析保持一致,考虑到学术文本的摘要和正文在复杂度上可能表现不一致,本文将其分开为两个样本进行分析。

3 学术文本可读性分析

3.1 学术文本内部要素的可读性分析

学术文本不同学科之间可能因研究内容、表达方式不同而产生可读性差异,学术文本摘要部分需要对正文提纲挈领,语言使用也可能不同。本文选取计算机学科(CS)和图书情报学科(LIS)的英文期刊论文,将摘要(ABS)与正文(PASS)分开,用SMOG、FKG、ARI、FOG 4个可读性指标进行计算,得出其平均值,结果如图1所示。摘要平均句子数量少于30,不适用SMOG公式。

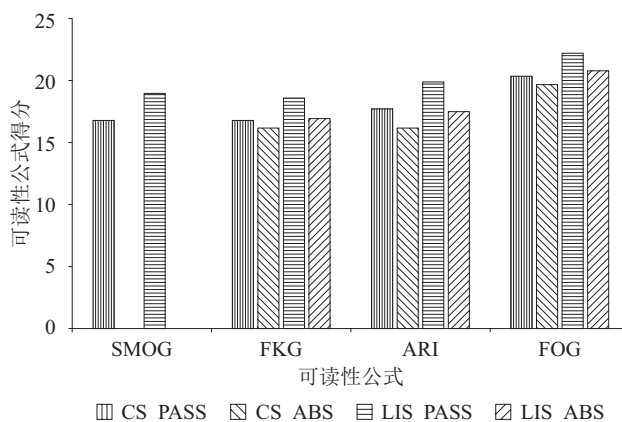


图1 学术文本不同学科、不同结构的可读性得分

从不同学科分析,LIS得分均明显高于CS。CS各公式得分平均值为17.72,LIS为19.31,差值为1.59。ARI公式得分差值最大,FKG最小。在不考虑文本内容的情况下,仅从语言层面分析,LIS学术文本比CS可读性差,即读者阅读时可能有更大的语言障碍,会遇到更多难词生词,更难以理解其中的长段落、长句子或短语等。

从不同结构上分析,两个学科的期刊论文正文文本得分都高于摘要。CS、LIS正文的可读性平均得分(后3个指标计算结果)为18.36和20.30,摘要为17.38和18.45,ARI公式得分差值最大,FOG最小。相对而言,CS与LIS相比,CS的正文和摘要得分差值更小。由

此可知,学术文本正文可读性不如摘要,在图书情报学科这一特点更加明显。

对于以上数值呈现的实际含义,本文仍然采用两独立样本 t 检验来验证。从学科比较来看,该检验的 F 统计量的观察值为0.001,对应的概率 p 值为0.970,大于0.05,可以认为两总体方差无显著差异。进一步考察 t 检验结果,发现 t 统计量的观测值为-1.697,对应的双尾概率 p 值为0.141,大于0.05,可知CS和LIS的可读性得分在平均值上有差异,但不显著。

从文本结构来看,CS、LIS检验的 F 统计量的观察值分别为0.379、0.734,对应的概率 p 值分别为0.538、0.397,均大于0.05,可以认为两总体方差无显著差异。进一步考察 t 检验的结果,发现 t 统计量的观测值分别为6.865、1.206,对应的双尾概率 p 值CS近似为0, LIS为0.236。如果显著性水平为0.05,由于CS概率 p 小于0.05,应该拒绝零假设,即CS英文期刊的正文和摘要可读性具有显著性差异;而LIS概率 p 大于0.05,不应该拒绝零假设,即LIS英文期刊的正文和摘要可读性虽有差异,但并不显著。

由于采用多个可读性公式进行计算,公式所得结果的一致性关系到对数据结果的分析是否准确合理。因此,本文用皮尔森相关系数评价选取的4个可读性公式的一致性。样本选取的是4个公式在113本CS期刊和6本LIS期刊中正文的可读性得分,共计119个。

由表2可知,各个公式之间的简单相关系数在0.7~1.0,说明4个可读性公式之间是正相关,对文本可读性评价方向上具有一致性,选取公式可信赖;相关系数检验的概率 p 值都近似为0,因此,当显著性水平为0.01时,都应拒绝零假设,认为它们总体上存在线性关系;相关系数大于0.8时,证明两者强相关,以上公式指标之间既有强相关关系,也有相对较弱关系,客观上正好反映本文所选取公式的全面性与合理性。

3.2 不同影响因子的期刊学术文本可读性分析

影响因子对学术文本非常重要,甚至成为衡量学术文本学术价值的重要标准。为比较不同影响因子期刊的学术文本可读性之间的差异,本文选取计算机学科的期刊论文,以期刊当年影响因子(Impact Factor, IF)和期刊5年影响因子(5 Year Impact Factor, 5 IF)为依据,将两者数值相加后排序,从113本CS期刊中分别选取排名靠前和靠后的10本期刊作为计算数据,相

表2 计算机学科和图书情报学科可读性公式的简单相关系数矩阵

	FKG	ARI	SMOG	FOG
FKG 皮尔森相关性	1.000	0.989	0.734	0.981
显著性(双尾)	-	0.000	0.000	0.000
ARI 皮尔森相关性	0.989	1.000	0.743	0.973
显著性(双尾)	0.000	-	0.000	0.000
SMOG 皮尔森相关性	0.734	0.743	1.000	0.810
显著性(双尾)	0.000	0.000	-	0.000
FOG 皮尔森相关性	0.981	0.973	0.810	1.000
显著性(双尾)	0.000	0.000	0.000	-

注:相关性在0.01上显著(双尾)。

关情况见表3。

图2展示了CS学科不同影响因子期刊学术论文的摘要、正文在不同可读性指标上的得分。

表3 不同影响因子的期刊学术文本

序号	类别	平均IF	平均5 IF	摘要数量	全文数据
A	高被引期刊	3.655 8	4.027 9	28 075	26 170
B	低被引期刊	0.669 7	0.761 8	14 543	7 078

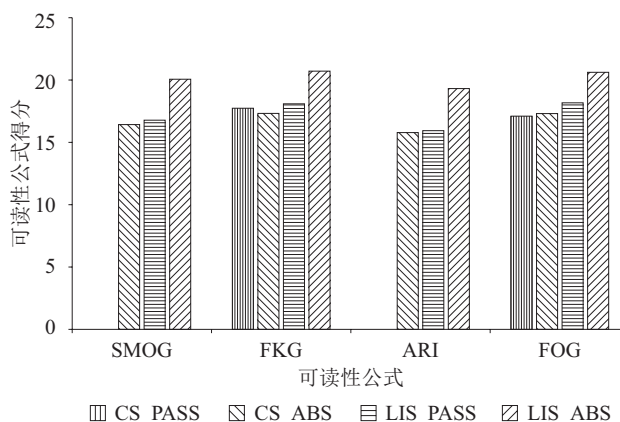


图2 CS学科不同影响因子期刊论文可读性得分

A类论文的得分均高于B类论文,其中摘要最为明显,平均相差0.7左右;但A类正文得分与B类差别很小,除SMOG的0.71外,其他情况下最多相差0.1,甚至在FKG上完全相等。用两独立样本 t 检验进行验证,发现两者双尾概率 p 值分别为0.670和0.864,都大于显著性水平0.05,说明两者差异并不明显。

因此,在不考虑研究内容、读者背景等情况下,仅从语言层面上进行分析,高被引期刊、低被引期刊的摘要、正文的可读性均不同,但从统计角度来看,这种差异并不明显。以CS期刊为例,高被引期刊刊载论文的摘要、正

文内容的可读性得分均高于低被引期刊的得分,即相较于低被引期刊而言,高被引期刊刊载论文对读者的语言能力要求稍高,且这种差异在摘要上体现得更加明显。

Gazni^[15]在研究文献摘要和可读性关系时,选取文章被引量最多的哈佛大学、斯坦福大学等5家机构的文献,发现文章被引和可读性间呈正相关关系,即文本越难以阅读,被引量越高,最高的相关系数甚至达到0.786。该趋势与本文结论一致,区别在于Gazni只使用了一个可读性公式,且在样本选择上,选择被引量最多的5家机构的22个学科文献,与本文的某一学科大量论文还略有区别。据此推测,可读性越低,被引越高的趋势,在高质量的文章中更明显。

4 学术文本句法复杂度分析

4.1 不同样本句法复杂度比较分析

经过L2SCA测量,得出不同学科学术文本句法复杂度的结果。

从平均数值上看,CS和LIS两个不同学科的对比,在摘要和正文两种结构上呈现出相反的趋势。LIS的正文得分有7项比CS高(如小句与句子比、小句与T单位比等),但是其摘要得分除T单位与句子比低0.002外,其余13项全部高于CS。由此可知,LIS论文正文比CS复杂度更高,但摘要反而更简单。

从文本内部结构来看,两个学科的正文与各自的摘要比较,正文得分基本高于摘要,只有并列短语与小句比、复杂名词性短语与小句比结果相反。因此,总体来看,学术文本正文比摘要句法复杂度更高,更倾向于使用长句和复杂语言。

4.2 学术文本句法复杂度指标双尾检验

为验证所得结果的准确度,本文也对所有指标进行双尾检验,结果见表4。

虽然同属一个学科,但由于摘要和正文承担的功能不同,因此在语言使用上可能依然存在差异性。

从双尾检验的概率 p 值看,只有计算机学科平均小句长的值大于0.05,即在这一指标上摘要和正文差异不显著,其他指标几乎接近于0,即都具有显著差异。具体从均值看,句子长度方面,正文比摘要长6~9个单位,这符合预期和实际情况,但在LIS学科,摘要在小句长

表4 计算机学科、图书情报学科学术文本
不同结构句法复杂度指标双尾检验结果

指 标	p 值(正文和摘要)		平均值(CS)		平均值(LIS)	
	CS	LIS	摘 要	正 文	摘 要	正 文
平均句长	0.000	0.000	23.584	32.898	24.575	31.274
平均T单位长	0.000	0.000	22.029	29.290	22.990	27.560
平均小句长	0.981	0.000	15.780	16.326	16.142	15.285
小句与句子比	0.000	0.000	1.554	2.036	1.593	2.075
小句与T单位比	0.000	0.000	1.444	1.809	1.478	1.824
复杂T单位与T单位比	0.000	0.000	0.363	0.449	0.380	0.472
从句与小句比	0.000	0.000	0.284	0.378	0.289	0.408
从句与T单位比	0.000	0.000	0.462	0.702	0.481	0.766
并列短语与小句比	0.000	0.002	0.445	0.408	0.592	0.547
并列短语与T单位比	0.002	0.000	0.616	0.723	0.824	0.964
T单位与句子比	0.000	0.000	1.078	1.124	1.079	1.137
复杂名词性短语 与小句比	0.000	0.000	2.345	2.311	2.384	2.115
复杂名词性短语 与T单位比	0.000	0.000	3.282	4.147	3.404	3.820
动词短语与T单位比	0.003	0.000	2.105	2.891	2.131	2.816

度方面大于正文。从句使用方面,正文使用比例大于摘要,在T单位上的差距最大。特定短语结构方面,CS学科复杂性名词短语在小句上几乎没有差距,LIS学科则摘要使用更多;在T单位上正文比摘要明显要大,正文使用更多。动词短语在T单位中出现的比例也是正文大于摘要。总体而言,两个学科的学术文本正文在大多数指标上都比摘要更复杂。

5 结语

本文综合利用FOG、SMOG、FKG、ARI4种可读性公式和句法复杂度测量工具L2SCA,从学科、文本结构和期刊影响力等方面对学术文本的可读性和句法复杂度进行探索。研究结果表明,不同学科学术文本间的可读性差异不明显,而正文部分的可读性显著低于摘要部分;高被引期刊的可读性得分高于低被引期刊,这一点在摘要上表现得更为明显;在句法复杂度方面,不同学科的学术文本句法复杂度有一定差别,且正文在大多数指标上都比摘要更复杂。

参考文献

[1] 李沂濛,赵良英,柯岚馨,等. 国内人文社会科学评价研究评析[J].

- 情报科学, 2017, 35 (7): 107-113.
- [2] 李信, 李旭晖, 陆伟, 等. 大数据驱动下的图书情报学科热点领域挖掘——面向WOS题录数据的实证视角[J]. 图书馆论坛, 2017, 37 (4): 49-57.
- [3] 李绍山. 易读性研究概述[J]. 解放军外国语学院学报, 2000, 23 (4): 1-5.
- [4] 陈洁. 国内可读性研究概述[J]. 黑龙江史志, 2013 (9): 212-213.
- [5] RICHARDS J C, PLATT F, PLATT H. 朗文语言教学及应用语言学辞典(英英·英汉双解)[M]. 北京: 外语教学与研究出版社, 2000.
- [6] ORTEGA L. Syntactic complexity measurement and its relationship with second language acquisition: a review of the studies of university second level writing[J]. Applied Linguistics, 2003, 24 (4): 492-518.
- [7] PERSSON T. Scientific language and readability: the correlation between the linguistic features of the TIMSS science project and the performance of different groups of grade 8 Swedish students[J]. Nordic Journal of Literacy Research, 2016 (2): 21-27.
- [8] ARMSTRONG J S. Unintelligible management research and academic prestige[J]. Interfaces, 1980, 10 (2): 80-86.
- [9] STREMERSCHE S, VERNIERI I, VERHOEF P C. The quest for citations: drivers of article impact[J]. Social Science Journal of Marketing, 2007, 71 (3): 171-193.
- [10] SAWYER A G, LARAN J, XU J. The readability of marketing journals: are award-winning articles better written?[J]. Journal of Marketing, 2014, 72 (1): 108-117.
- [11] DOLNICAR S, CHAPPLE A. The readability of articles in tourism journals[J]. Annals of Tourism Research, 2015, 52: 161-166.
- [12] SUN Y C. Does text readability matter? A study of paraphrasing and plagiarism in English as a foreign language writing context[J]. Asia-Pacific Education Researcher, 2012, 21 (2): 296-306.
- [13] 方龙, 李信, 黄永, 等. 学术文本的结构功能识别——在关键词自动抽取中的应用[J]. 情报学报, 2017, 36 (6): 599-605.
- [14] PLAVÉN-SIGRAY P, MATHESON G J, SCHIFFLER B C, et al. The readability of scientific texts is decreasing over time[EB/OL]. (2017-04-28) [2018-04-03]. <https://doi.org/10.1101/119370>.
- [15] GAZNI A. Are the abstracts of high impact articles more readable? Investigating the evidence from top research institutions in the world[J]. Journal of Information Science, 2011, 37 (3): 273-281.
- [16] LEI L, YAN S. Readability and citations in information science: evidence from abstracts and articles of four journals (2003—2012)[J]. Scientometrics, 2016, 108 (3): 1-15.
- [17] 陆小飞, 许琪. 二语句法复杂度分析器及其在二语写作研究中的应用[J]. 外语教学与研究, 2016, 48 (3): 409-420.
- [18] 索传军, 盖双双, 周志超. 认知计算——单篇学术论文评价的新视角[J]. 中国图书馆学报, 2018, 44 (1): 50-61.
- [19] 李力, 刘德洪, 张灿影. 基于知识流动理论的科技论文学术影响力评价研究[J]. 情报科学, 2016, V34 (7): 113-119.
- [20] 徐峰, 冷伏海. 认知计算及其对情报科学的影响[J]. 情报杂志, 2009, 28 (6): 20-23.

作者简介

陈练文, 女, 1976年生, 博士, 讲师, 研究方向: 文本可读性、语言计量, E-mail: chenlianwen@126.com。

李信, 男, 1991年生, 博士研究生, 研究方向: 文本挖掘、信息计量, E-mail: lucian@whu.edu.cn。

赵超辉, 男, 1991年生, 硕士研究生, 研究方向: 文本可读性、信息计量。

Analysis of the Readability and Complexity of Academic Texts

CHEN LianWen¹ LI Xin² ZHAO ChaoYe²

(1. College of Chinese Language and Literature, National Institute of Chinese Language Matters and Social Development, Wuhan University, Wuhan 430072, China;

2. School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: This article chooses academic texts from computer science and library and information science as research objects, using four classical readability formulas SMOG, FKG, ARI and FOG and syntactic complexity tool L2SCA, to analyze the readability and syntactic complexity of academic texts from research domain, text structures, and journal impact factor perspectives. The results show that texts from different research domain has no difference on readability which is not consistent to different part of academic texts.

Keywords: Academic Text; Text Readability; Syntactic Complexity; Readability Formula; Complexity Evaluation

(收稿日期: 2018-05-03)