

柯尔莫哥洛夫复杂度:理论、方法、应用与展望

肖巍^{1,2} 刘洵¹

(1. 重庆大学 外国语学院,重庆 401331;2. 重庆大学 语言认知及语言应用研究基地,重庆 401331)

摘要:柯尔莫哥洛夫复杂度是一种基于信息论的语言复杂度测量体系,可测量形态、句法和整体层面的复杂度,具有客观性、经济性、真实性和全面性的特点。然而,国内语言学界对其引介尚不充分。本文首先介绍了柯尔莫哥洛夫复杂度的理论基础,以及形态复杂度、句法复杂度和整体复杂度的计算方法;随后,回顾了柯尔莫哥洛夫复杂度在语言研究当中的实证应用,发现早期研究主要涉及语言类型学,近期也开始关注语域变体,并从本体拓展到二语;最后,展望了柯尔莫哥洛夫复杂度在丰富语义语用测量指标、拓展更多应用领域、深化汉语本土研究等方面的未来发展。本文希冀加深国内学者对柯尔莫哥洛夫复杂度的认识,丰富国内学者探究语言复杂度的测量手段,推动汉语语言学、汉外语言对比、国际中文教育等研究领域的发展。

关键词:Kolmogorov 复杂度;语言复杂度;测量;信息论

中图分类号:H0-06 **文献标志码:**A **文章编号:**1674-6414(2025)02-0018-13

0 引言

语言复杂度(language complexity)指语言系统中各成分的数量、种类和交互关系的复杂程度,以及语言使用者对各成分的认知难度(Hulstijn et al., 1994; Bulté et al., 2012; Pallotti, 2015)。对语言复杂度进行探究,既可助力语言结构描写、对比语言学、语言类型学等本体研究,也可应用于语言能力测评、语言能力发展、第二语言教学等领域。在语言复杂度研究中,一个核心问题即如何精准有效地测量语言复杂度。目前,学者业已从语言的不同层面,包括词汇(Ure, 1971; Linnarud, 1986; Lu, 2012; 鲍贵, 2008)、句法(Hunt, 1965; 鲍贵, 2009)、形态(Horst et al., 2006; Brezina et al., 2019)、短语搭配(Church et al., 1990; Paquot, 2019; 胡韧奋, 2021)等,探讨了语言复杂度的测量问题。基于这些探究,学

收稿日期:2024-03-17

基金项目:重庆市社会科学规划培育项目“复杂动态系统视域下来华留学生汉语写作复杂度发展模式研究”(2024PY17)和教育部人文社会科学研究青年基金项目“基于文本挖掘的中美哲学社会科学学术话语对比研究”(23XJC740006)的阶段性成果

作者简介:肖巍,男,重庆大学外国语学院教授,博士,博士生导师,语言认知及语言应用研究基地研究员,主要从事计量语言学、语言大数据挖掘和学术话语研究。

刘洵,男,重庆大学外国语学院硕士研究生,主要从事计量语言学研究。

引用格式:肖巍,刘洵. 柯尔莫哥洛夫复杂度:理论、方法、应用与展望[J]. 外国语文, 2025(2): 18-30.

界提出了一系列语言复杂度的测量指标,如词汇密度(lexical density)(Ure, 1971)、T 单位长度(LTU)(Hunt, 1965)、形容词修饰语词根类型标记比率(amod_RTTR)(Paquot, 2019)等。这些测量指标被广泛应用于语言能力测量(Kim et al., 2017; 雷蕾, 2017)、语言对比(Nichols, 1992; 陈蕊娜, 2017)、语域变体(Xiao et al., 2020)等话题,为相关领域的研究提供了精准的定量支撑。

诚然,语言复杂度的测量方法已得到长足的发展。然而,现有的测量手段仍存在若干提升空间。首先,现有方法大都聚焦于语言微观层面的“局部复杂度”(local complexity)(Miestamo, 2009: 29-32)。在测量宏观层面上的“全局复杂度”(global complexity)时,研究者需汇集多个层面的微观指标构建庞杂的测量指标体系,可能导致指标间的高共线性与高冗余度,在实践应用时也容易造成指标选取与解读上的困难。其次,大多数测量方法需要进行词性等特征的识别与标注,这一过程可能出现标注错误或掺杂标注者的主观认知,从而导致计算结果的偏误。再次,囿于语间差异,一些基于特定语言特征的测量方法并不适用于不存在该特征的语言,其跨语言对比研究的应用受到一定限制。

为解决上述问题,有学者开始尝试探索更加注重全局性、客观性与普适性的测量方法。其中,一种基于信息论的语言复杂度测量体系——柯尔莫哥洛夫复杂度(Kolmogorov complexity)——逐渐受到语言学的关注。该方法最早由苏联数学家安德雷·柯尔莫哥洛夫(Andrey Kolmogorov)(1965)提出,用于测量一串信息的“最短可能描述长度”(Li et al., 2004: 3252)。在语言学界,帕特里克·乔拉(Patrick Juola)(1998/2008)首次将柯尔莫哥洛夫复杂度引入语言复杂度的研究中。此后,凯瑟琳娜·埃雷特(Katharina Ehret)与本尼迪克特·斯姆雷奇萨尼(Benedikt Szmrecsanyi)(2016)将其拓展为一套覆盖形态、句法、整体三个层面的系统性的语言复杂度测量体系。在国外,柯尔莫哥洛夫复杂度,尤其是基于埃雷特-斯姆雷奇萨尼(Ehret-Szmrecsanyi)测量体系的柯尔莫哥洛夫复杂度,已取得较为充分的发展和应用。然而,国内语言学界对其引介尚不充分,相关实证应用更是寥寥。为此,本文将对柯尔莫哥洛夫复杂度的理论基础、计算方法与实证应用进行评述,展望其发展前景,希冀加深国内语言学界对柯尔莫哥洛夫复杂度的认识,丰富国内学者探究语言复杂度的测量手段,推动汉语语言学、汉外语言对比和国际中文教育等研究领域的发展。

1 柯尔莫哥洛夫复杂度的理论基础

柯尔莫哥洛夫复杂度描述一个相对“简单”的对象,所需表述较短,即需要较少的信息;反之,描述一个相对“复杂”的对象,所需表述较长,即需要较多的信息(吴彤等, 2003a/2003b)。若一串信息的“最短可能描述长度”较短,则其信息量较小,反之信息量较大。例如:

(1) cdgh39aby7

(2) cdcdcdcdcd (Ehret et al., 2016: 75)

两例均由 10 个字符组成。观察可知,例(1)不存在重复规律,而例(2)存在明显的重复规律,即字符“cd”重复了五次。若在不损失信息的前提下描述这两串字符,例(1)只能以原字符描述,而例(2)可以压缩描述为“5×cd”。“cdgh39aby7”长于“5×cd”,故例(1)的“最短可能描述长度”比例(2)要长,即例(1)比例(2)承载了更多的信息,其在柯尔莫哥洛夫复杂度意义上更加复杂。

柯尔莫哥洛夫复杂度可通过勒姆佩尔-齐夫(Lempel-Ziv)压缩算法进行测量(Li et al., 1997; Ziv et al., 1977)。测量时,该算法预先在缓存中载入一定信息片段。在遍历后续信息时,若后续信息已存在于缓存,就可对其进行压缩(Ehret, 2014: 47)。压缩后的信息可视为该信息的“最短可能描述长度”,其大小即为该信息柯尔莫哥洛夫复杂度的值。

在语言学领域,美国计算语言学家乔拉(1998)首次引入柯尔莫哥洛夫复杂度进行语言复杂度测量(Szmrecsanyi, 2015)。其基本测量逻辑为:能够被更好、更有效压缩的文本,在语言上相对更简单;反之则更复杂(Juola, 1998/2008)。基于该逻辑,乔拉(1998)提出了基于柯尔莫哥洛夫复杂度的“退化”(degradation)方法,用于测量文本的形态复杂度。此后,乔拉(2008)又提出了“扰乱”(distortion)方法。该方法既可用于形态层面,也可用于句法层面的复杂度测量。乔拉的系列研究奠定了柯尔莫哥洛夫复杂度的语言学理论基础。

乔拉的方法主要聚焦语言形态、句法等微观层面,并未着眼于宏观层面上的“全局复杂度”(Miestamo, 2009: 29-32)。为解决这一问题,埃雷特与斯姆雷奇萨尼吸收并拓展了乔拉提出的方法,形成了一套较为完善的柯尔莫哥洛夫语言复杂度测量体系(Ehret, 2014/2016/2018)。在该体系中,柯尔莫哥洛夫复杂度被定义为测量“结构表层冗余”(structure surface redundancy)的指标(Ehret, 2014: 47)。其中,“结构表层”指该测量体系主要基于语言结构形式而非语义功能进行计算(Ehret et al., 2016: 75),“冗余”指语言中“没有显性语法功能的语言标记、形式或类别”(McWhorter, 2001: 129),从测量角度看则是文本中字符或结构的复现(Ehret, 2016)。柯尔莫哥洛夫复杂度测量体系基于“文本中普遍存在冗余成分”的假设,运用基于勒姆佩尔-齐夫算法的压缩程序识别出语言结构表层的字符、词汇或结构冗余并进行压缩,从而对语言的形态、句法及整体复杂度进行测量。可见,柯尔莫哥洛夫复杂度虽然理论基础源远流长,但在语言学界直到近十年才发展为一套系统完备的语言复杂度测量体系。

相比其他方法,柯尔莫哥洛夫复杂度具备四个特点(Ehret, 2016: 3-4):

(1) 客观性:柯尔莫哥洛夫复杂度无需事先对语言特征进行主观选择或主观判定语言特征的复杂程度;

(2) 经济性:柯尔莫哥洛夫复杂度的测量方法简单通用,理论上适用于任何语言、任何类型的语料库;

(3) 真实性:柯尔莫哥洛夫复杂度分析自然、真实的语言样本,测量实际使用中的语言复杂度。

(4) 全面性:柯尔莫哥洛夫复杂度不仅涵盖句法、形态等微观层面,还可以测量宏观层面上的整体复杂度。

2 柯尔莫哥洛夫复杂度的计算方法

2.1 形态复杂度的计算方法

相比其他语言层面,形态复杂与否更加直观,是一个理想的“试验台”(Juola, 1998: 207)。形态复杂度的计算方法主要有“退化”(degradation)(Juola, 1998)和“扰乱”(distortion)(Juola, 2008; Ehret, 2016)两种。

“退化”即将文本中的每种词形替换为随机符号(如数字),再对原文本和退化文本进行压缩,获取两个压缩文本的大小,计算二者比值作为形态复杂度的计量指标(Morphological Complexity, MC)。其计算公式如下:

$$MC = \frac{R}{C}$$

其中, R 为原文本压缩后的大小, C 为退化文本压缩后的大小。 MC 的数值越大,文本在形态上越复杂;反之越简单。退化方法可通过表1的示例加以演示。

表1 退化处理示例(Juola, 1998: 209)

原单词			退化单词		
Jump	Walk	Touch	8 634	139	5 543
Jumped	Walked	Touched	15	4 597	1 641
Jumping	Walking	Touching	3 978	102	6

在表1中,原文本单词之间(如 jump/jumped/jumping, jumped/walked/touched)存在形态变化规律,故有一定的压缩空间。退化操作将这些单词以随机数字替代(如 Touch 替换为 5 543, Touched 替换为 1 641),导致单词之间原本的形态变化规律消失,压缩空间变小。因此,无论原文本形态复杂与否,退化文本都难以压缩, C 都较大,可以近似视作一个常量。这样一来, MC 大小的关键就取决于 R 的大小。若原文本形态简单,就能够达到较高的压缩率,使得 R 相对较小, R 和 C 的比值 MC 也较小;若原文本形态复杂,就难以压缩,使得 R 相对较大, R 和 C 的比值 MC 也较大。

“扰乱”即随机删除文本中一定比例的字符(通常为 10%)(Juola, 1998),再对原文本

和扰乱文本进行压缩,计算两个压缩文本的大小,将二者比值的相反数定义为形态复杂度分数(Morphological Complexity Score, MCS)。其计算公式如下:

$$MCS = -\frac{m}{c}$$

其中, m 为扰乱文本压缩后的大小, c 为原文本压缩后的大小,比值取相反数是为了使 MCS 的数值大小与形态复杂程度呈正相关,以符合人类直觉。 MCS 越大,语言形态越复杂;反之越简单。为保证统计学上的稳健性,扰乱通常会重复若干次,取重复后的 MCS 平均值代表该文本的形态复杂度。

扰乱方法可通过下例进行演示:

(3) That day is a beautiful day.

(4) That dy is a beautiful day.

例(4)是例(3)形态扰乱后的结果,第二个单词“day”中的字母“a”被删除。在例(3)中,“day”重复出现两次,存在一定压缩空间。而例(4)中,“dy”与“day”呈现两种不同的词形态,并非完全重复,故无法同例(3)一样进行压缩。可见,无论原文本形态复杂与否,扰乱文本都难以压缩, m 都较大,可以近似视作一个常量。这样一来, MCS 大小的关键就取决于 c 的大小。若原文本单词种类与形式较少(即“形态简单”),就能够达到较高的压缩率,使得 c 相对较小, m 和 c 比值的相反数 MCS 也较小;若原文本单词种类与形式较多(即“形态复杂”),就难以压缩,使得 c 相对较大, m 和 c 比值的相反数 MCS 也较大。需要指出的是: MCS 计算公式中添加负号,是为了使数值大小与复杂度大小成正比关系;但负号的出现将数值变成了负数,在测量时略不符合人类直觉。因此,采用 c 和 m 的比值,即 $\frac{c}{m}$ 作为形态复杂度分数,可能是一个更好的办法。

2.2 句法复杂度的计算方法

句法作为语言系统的核心架构,其复杂度亦受到重点关注。柯尔莫哥洛夫句法复杂度的计算方法主要有“修改”(modification)(Sadeniemi et al., 2008)和“扰乱”(distortion)(Juola, 2008; Ehret, 2016)两种。

“修改”即将文本中每个句子的词序打乱,再对原文本和修改文本进行压缩,获取两个压缩文本的大小,计算二者的比值作为句法复杂度的计量指标(Syntactic Complexity, SC)。其计算公式如下:

$$SC = \frac{R}{M}$$

其中, R 为原文本压缩后的大小, M 为修改文本压缩后的大小。 SC 的数值越大,句法越复杂;反之越简单。

修改方法可通过下例进行演示:

(5) That day is a beautiful day.

(6) Beautiful that a is day day.

例(6)为例(5)进行修改后的结果。在例(5)中,词与词之间存在一定的词序和句法关系。根据这些句法规律,可一定程度对句子成分进行压缩。在例(6)中,词序被打乱,改变了正常的句法结构,故无法同例(5)一样进行压缩。可见,无论原文本句法复杂与否,修改文本都难以压缩, M 都较大,可以近似视作一个常量。若原文本句法严格、词序固定,就能够达到较高的压缩率,使得 R 相对较小, R 和 M 的比值 SC 也较小;若原文本句法灵活、词序自由,就难以压缩,使得 R 相对较大, R 和 M 的比值 SC 也较大。

“扰乱”即随机删除文本中一定比例的词(通常为10%),再对原文本和扰乱文本进行压缩,计算两个压缩文本的大小,将二者的比值定义为句法复杂度分数(Syntactic Complexity Score, SCS)。其计算公式如下:

$$SCS = \frac{s}{c}$$

其中,“ s ”为扰乱文本压缩后的大小,“ c ”为原文本压缩后的大小。 SCS 的数值越大,句法越复杂;反之越简单。为保证统计学上的稳健性,扰乱通常会重复若干次,取重复后的 SCS 平均值代表该文本的句法复杂度。

扰乱方法可通过下例进行演示:

(7) That day is a beautiful day.

(8) That is a beautiful day.

例(8)是例(7)句法扰乱后的结果。在例(7)中,词与词之间存在一定的词序和句法关系。根据这些句法信息,可一定程度对句子成分进行压缩。在例(8)中,第二个单词 *day* 被删除,改变了句法结构([NP [D that] [N day]]的中心语[N day]缺失),无法同例(7)一样进行压缩。可见,无论原文本句法复杂与否,扰乱文本都难以压缩, s 都较大,可以近似视作一个常量。若原文本句法严格、语序固定,就能够达到较高的压缩率,使得 c 相对较小, s 和 c 的比值 SCS 较大;若原文本句法灵活、词序自由,就难以压缩,使得 c 相对较大, s 和 c 的比值 SCS 较小。

2.3 整体复杂度的计算方法

相比微观层面的形态和句法复杂度,整体复杂度可以测量与解释文本宏观层面的“整体结构复杂度”(Ehret et al., 2021)。其计算方法为:获取原文本压缩前与压缩后的大小,以压缩前的文本大小为自变量,以压缩后的文本大小为因变量,进行线性回归拟合,得到回归模型:

$$y = ax + b + \varepsilon$$

其中, x 为压缩前文本大小, y 为压缩后文本大小, a 为斜率, b 为截距, ε 为回归残差。 ε 被定义为“校正的整体复杂度分数”(Adjusted Overall Complexity Score, AOCS),作为柯尔莫哥洛夫整体复杂度的计量指标。

整体复杂度的计算思路与形态和句法复杂度有所不同。形态或句法复杂度的计算思路为:压缩算法能够借助形态或句法规律进行文本压缩;如果在形态或句法层面进行一定的处理,就可以干扰原本的形态或句法规律,影响文本压缩率;形态/句法处理对文本压缩率的影响大小即可视作形态/句法复杂度的计量指标。整体复杂度并不聚焦于某一层面因素的影响,而是关注文本在整体上难以压缩部分的比重。这一比重可借助回归模型加以提取。在统计学上,回归模型可以预测自变量对因变量的影响,残差即因变量无法由自变量解释和预测的部分。若以压缩前的文本大小为自变量预测压缩后的文本大小,残差的大小则可理解为文本整体上不存在明显规律(难以预测和压缩)部分的比重,即整体复杂度。残差越大,难以预测和压缩的部分就越多,语言整体上就越复杂;反之则越简单。

3 柯尔莫哥洛夫复杂度的实证应用

自柯尔莫哥洛夫复杂度引入语言学界,尤其是埃雷特-斯姆雷奇萨尼测量体系推出以来,不少学者开始关注这一新方法,并尝试将其投入到具体的应用之中。相关实证应用早期主要涉及语言类型学,近期亦开始关注语域变体,并有从语言本体研究拓展到二语研究的趋势。

3.1 语言类型学

柯尔莫哥洛夫复杂度不依赖具体语言特征的特点,可为“所有语言是否同样复杂”的类型学争议提供数据支撑,也可为不同语言的谱系归类提供定量依据。探讨“所有语言是否同样复杂”,通常将多种语言纳入对比,并对复杂程度进行排序(Juola, 1998; Juola, 2008; Sadeniemi et al., 2008)。例如,乔拉(Juola, 1998/2008)分别使用退化和扰乱两种方法,以《圣经》为语料对比了不同语言的复杂度,发现:在退化方法下,芬兰语的形态复杂度最高,毛利语最低。在扰乱方法下,形态方面,印尼语最简单,乌克兰语最复杂,而句法方面,海地克里奥尔语最复杂,芬兰语最简单。埃雷特等(2016)运用形态及句法扰乱方法和整体复杂度计算方法对平行、半平行的文学作品语料及非平行新闻语料库进行测量,得出了与乔拉(Juola, 1998/2008)相似的结果,即芬兰语、匈牙利语等表现出较高的整体、形态复杂度与较低的句法复杂度,而英语、法语等与之相反。基莫·凯图宁(Kimmo Kettunen)等人(2006)与马库斯·萨德涅米(Markus Sadeniemi)等人(2008)则以《欧洲宪法》为语料,运用退化和修改方法对欧盟各国官方语言的形态、句法与形态句法复杂度进行计算与

对比,发现形态方面,德语最复杂,意大利语最简单;句法方面,芬兰语最复杂,法语最简单。马塞洛·蒙特穆罗(Marcelo A. Montemurro)与达米安·扎内特(Damián H. Zanette)(2011)以八种语言的非平行文本为语料,利用勒姆佩尔-齐夫压缩算法,探索人类语言在词序上是否存在普遍机制。结果发现尽管不同语言在结构和词汇上存在差异,但词序对语言结构的影响具有统计学上的语言普遍性。上述研究使用柯尔莫哥洛夫复杂度分析了不同语言、不同类型的语料,为类型学议题提供了精准的数据支撑,并极大提升了类型学研究的科学性和可复现性。

对不同语言进行谱系归类,主要以柯尔莫哥洛夫复杂度为特征,通过无监督/监督学习进行自动归类(Li et al., 2004; Cilibrasi et al., 2005)。例如,李明(Ming Li)等人(2004)以《世界人权宣言》为平行语料库,构建了52种亚欧语言的语言谱系,将不同语言较为准确地归类到其语支中,并从一些错误的归类(如将英语归类到罗曼语支)中捕捉到不同语言间相互交流与融合的证据。鲁迪·西利布拉西(Rudi Cilibrasi)与保罗·维塔尼(Paul Vitányi)(2005)在李明等人(2004)的基础上加以改进,成功进行了更精确的语支自动归类。这些基于柯尔莫哥洛夫复杂度的谱系归类取得成功,从信息论角度揭示了不同语言之间的共性与个性,表明柯尔莫哥洛夫复杂度可能是语言符号在信息层面更为概化和稳健的内在属性和计量指标。

3.2 语域变体

在探究不同语言的复杂度之外,同一语言的语域变体也是研究者关注的领域。该领域的研究主要基于大规模语料库,对比不同语域变体的语言复杂度(Ehret, 2018)、探究语域变体复杂度与其他语言因素的关系(Ehret et al., 2021)及语域变体复杂度的历时变化(Wang et al., 2022a)等。例如,埃雷特(Ehret, 2018)计算了来自英国国家语料库(BNC)的4 053篇文本的柯尔莫哥洛夫复杂度,用以对比英语口语和书面语域、正式与非正式语域的复杂度异同。结果发现:书面、正式语域(如学术语篇)在整体和形态上比口语、非正式语域(如口语对话)更加复杂,在句法上则比后者简单;随着语体正式程度增加,文本呈现出整体与形态复杂度不断提高、句法复杂度不断降低的趋势。埃雷特等(Ehret et al., 2021)基于西蒙菲莎大学意见与评论语料库(SOCC),探究了新闻与意见型话语(包括评论文章与网络评论)中语言复杂度与语言主观性的关系。结果表明,新闻相较意见型话语在柯尔莫哥洛夫复杂度上更加复杂,两者的主观性标记数量并没有显著区别,说明较客观的语域在语言复杂度上与较主观语域存在明显区别;但无论语域客观与否,都会在一定程度上传达语言使用者的主观意见、立场或评价。王贵(Gui Wang)等人(2022a)以《伦敦皇家学会哲学学报》1820年至1920年间的文章为语料,探讨了英语科学写作语言复杂度的历时变化。研究发现,100年间,科学写作的全文整体与形态复杂度呈现上升趋势,句法复杂度呈

现下降趋势,说明英语科学写作所用词汇规模大幅扩张,而句法则逐渐简化;该研究还发现了句法和形态复杂度之间的权衡效应,以及摘要部分的复杂度与全文呈现完全相反的发展趋势。

上述研究为语域变体领域提供了新方法,开拓了新思路,并提出了语域分类、假新闻识别等新的应用场景,表明柯尔莫哥洛夫复杂度作为语言符号,在信息层面更为概化和稳健的内在属性和计量指标可以应用到更多的研究领域和语境。

3.3 二语研究

近几年来,柯尔莫哥洛夫复杂度的应用开始从语言本体拓展到二语研究。该领域的研究主要致力于探讨二语写作的语言复杂度与其他因素间的关系(Ehret et al., 2019; Ströbel et al., 2020; Wang et al., 2022b)。例如,埃雷特与斯姆雷奇萨尼(2019)通过对国际英语学习者语料库(ICLE)中不同水平英语学习者论文写作进行伪纵向研究,探讨英语二语论文的语言复杂度与论文作者英语学习时长之间的关系。其结果表明,随着学习时间增加,英语二语论文的柯尔莫哥洛夫复杂度在整体与形态层面持续提高,而在句法层面不断降低;此外,对于英语学习时间较长(10年以上)的学习者,其论文的整体与形态复杂度甚至超越了本族语者。马库斯·斯特罗贝尔(Marcus Ströbel)等人(2020)基于柯尔莫哥洛夫复杂度等测量指标,对80名德语母语者的英语二语写作与德语写作用本进行测量,探讨二语写作复杂度与一语写作复杂度之间的关系。研究发现,二语写作与一语写作的柯尔莫哥洛夫复杂度呈现较强的正相关。即使在控制年龄与英语国家经历两个背景变量后,该正相关关系仍然存在,表明一语写作能力较强的个体在二语写作中也会表现出较高的语言复杂度,一语语言能力应被纳入二语学习者个体差异研究的考量因素当中。王贵等人(2022a)基于774篇中国英语学习者的议论文语料,将柯尔莫哥洛夫复杂度与传统的词汇、形态与句法复杂度指标进行了对比,发现柯尔莫哥洛夫复杂度相比于传统指标,在整体、句法层面能够更好地区分不同二语学习水平,但柯尔莫哥洛夫句法复杂度与一些基于依存关系的细粒度复杂度指标(如名词主语中依存成分均数、从句中名词主语均数)之间不存在显著相关。作者据此建议结合柯尔莫哥洛夫复杂度与依存关系细粒度指标,以提高二语语言复杂度测量的全面性和准确性。

4 未来展望

柯尔莫哥洛夫复杂度自引入语言学研究以来,理论基础不断夯实,从微观层面拓展到宏观层面;计算方法逐渐完善,形成了成熟的测量体系;实证应用日趋丰富,研究话题涵盖多个领域。未来,柯尔莫哥洛夫复杂度研究有望在如下方面进一步发展。

理论方面,未来研究可继续加深柯尔莫哥洛夫复杂度的测量层面,开发更多的测量指

标。目前的测量层面包括形态、句法和整体,但语义、语用等层面的复杂度还鲜有涉及。乔拉(2008)曾提出“语用复杂度”的概念,以“删除一定比例的句子扰乱语篇”的方法测量语篇层面的复杂度,但后续研究未能顺此脉络及时跟进。若能开发有效的语义和语用复杂度测量指标,必将极大推动语义学、语用学、话语分析等相关领域的研究。

实证方面,可继续拓展柯尔莫哥洛夫复杂度在更多语言研究领域的应用。柯尔莫哥洛夫复杂度指标体系简明,可以更加直观、综合地展示各层面的语言复杂度特征;不依赖具体语言特征,可以应用于不同语言的复杂度测量当中,具有较强的通用性;测量简单易行,实现难度较低,利用已发布的脚本^①(Ehret, 2018: 6)便可实现自动测量。基于上述优点,柯尔莫哥洛夫复杂度在诸多领域都具有广阔的应用前景。例如,在语言对比领域,柯尔莫哥洛夫复杂度可作为不同语言对比的特征指标之一,通过挖掘其反映的语言特征,为相关理论提供定量证据;在阅读研究领域,可利用柯尔莫哥洛夫复杂度指标进行多层面的可读性评估,为教师提高阅读教学效果、选择合适的阅读材料提供帮助;在二语习得与测评领域,可结合动态复杂系统理论,探究二语能力的动态发展、停滞与倒退,并可利用柯尔莫哥洛夫复杂度指标监测学生的二语学习效果,及时反馈,以评促学。

本土化方面,未来研究可积极开展柯尔莫哥洛夫复杂度在汉语研究中的理论探索与实践应用。当前,相关研究多基于西方语言,对汉语的关注不多。研究汉语的柯尔莫哥洛夫复杂度,既能够将外国的理论方法本土化,解决中国的实际问题,又能够将本土的研究成果国际化,用中国特色、中国风格、中国气派的语言学研究引领国际语言学界的学术前沿。目前已有学者注意到这一研究领域,业已验证了柯尔莫哥洛夫复杂度在汉语语言复杂度测量中的信度与效度(Liu et al., 2024),开发了基于柯尔莫哥洛夫复杂度的汉语词汇结构复杂度测量技术(Hong et al., 2023)。然而,如何针对汉语与汉字的特殊性,设计一整套测量汉语各个层面柯尔莫哥洛夫复杂度的测量体系,并应用于汉语语言学、汉外语言对比、国际中文教育等领域,尚需更加深入的探索,也是我们中国语言学人“扎根本土、面向国际”所应进一步思考并解决的问题。

5 结语

本文引介了一种源自信息论的语言复杂度测量体系——柯尔莫哥洛夫复杂度,介绍了其概念、提出及在语言学领域的引入与发展,阐释了其理论基础与计算方法,梳理回顾了相关实证应用,并对其发展前景进行了展望。经过数十年发展,柯尔莫哥洛夫复杂度已经形成一套较为完善的测量体系,具有客观、经济、真实、全面等特点,能够展现形态、句法、整体三个层面的语言复杂度特征,并已应用于语言类型学、语域变体、二语研究等领域,在理论、

^① 脚本网址:<https://github.com/katehret/measuring-language-complexity>. [2024-01-15].

实证和本土化方面均有着广阔的应用前景。

诚然,柯尔莫哥洛夫复杂度也存在一些局限。理论上,由于其压缩算法对语言的形式-功能配对结构不可知,因而对细粒度的语言特征欠敏感,着力于宏观测量而乏力于微观分析。应用上,虽然现有测量工具已经较为便利,但对于欠缺技术背景的使用者来说,仍存在一定的使用门槛。未来可以开发网页版应用、桌面应用程序、手机客户端、小程序等扩展工具,以增强测量的便利性和易用性。尽管如此,考虑到其鲜明优点和广阔前景,仍然值得进一步推广应用。

当前,柯尔莫哥洛夫复杂度在国际语言学界已取得较为充分的发展和应用,但在国内的引介尚不多见。通过本文引介,希冀加深国内语言学界对柯尔莫哥洛夫复杂度的认识,丰富国内学者探究语言复杂度的测量手段,推动汉语语言学、汉外语言对比、国际中文教育等研究领域的发展,为国外理论方法的本土化与中国语言学的国际化贡献绵薄之力。

参考文献:

- Brezina, V. & G. Pallotti. 2019. Morphological Complexity in Written L2 Texts [J]. *Second Language Research* 35 (1): 99-119.
- Bulté, B. & A. Housen. 2012. Defining and Operationalising L2 Complexity [G] // A. Housen, I. Vedder & F. Kuiken. *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Amsterdam: John Benjamins B. V., 23-46.
- Church, K. & P. Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography [J]. *Computational Linguistics* (1): 22-29.
- Cilibrasi, R. & P. M. Vitányi. 2005. Clustering by Compression [J]. *IEEE Transactions on Information Theory* (4): 1523-1545.
- Ehret, K. 2014. Kolmogorov Complexity of Morphs and Constructions in English [J]. *Linguistic Issues in Language Technology* (2): 43-71.
- Ehret, K. 2016. *An Information-Theoretic Approach to Language Complexity: Variation in Naturalistic Corpora* [D]. Freiburg: Albert-Ludwigs-Universität Freiburg.
- Ehret, K. 2018. An Information-Theoretic View on Language Complexity and Register Variation: Compressing Naturalistic Corpus Data [J]. *Corpus Linguistics and Linguistic Theory* (2): 383-410.
- Ehret, K. & B. Szmrecsanyi. 2016. An Information-theoretic Approach to Assess Linguistic Complexity [G] // R. Baechler & G. Seiler. *Complexity, Isolation, and Variation*. Berlin: de Gruyter, 57-71.
- Ehret, K. & B. Szmrecsanyi. 2019. Compressing Learner Language: An Information-theoretic Measure of Complexity in SLA Production Data [J]. *Second Language Research* (1): 23-45.
- Ehret, K. & M. Taboada. 2021. The Interplay of Complexity and Subjectivity in Opinionated Discourse [J]. *Discourse Studies* (2): 141-165.
- Hong, X., Huang, W. & H. Liu. 2023. The Structural Complexity of Chinese Words and Its Relationship with Word Frequency

- [J]. *Journal of Quantitative Linguistics* (3-4): 231-256.
- Horst, M. & L. Collins. 2006. From Faible to Strong; How Does Their Vocabulary Grow? [J]. *Canadian Modern Language Review* (1): 83-106.
- Hulstijn, J. H. & R. De Graaff. 1994. Under What Conditions Does Explicit Knowledge of a Second Language Facilitate the Acquisition of Implicit Knowledge? A Research Proposal [J]. *AILA Review* (11): 97-112.
- Hunt, K. 1965. *Grammatical Structures Written at Three Grade Levels* [R]. NCTE Research Report No 3.
- Juola, P. 1998. Measuring Linguistic Complexity: The Morphological Tier [J]. *Journal of Quantitative Linguistics* (3): 206-213.
- Juola, P. 2008. Assessing Linguistic Complexity [G] // M. Miestamo, K. Sinnemki & F. Karlsson. *Language Complexity: Typology, Contact, Change*. Amsterdam, Philadelphia: Benjamins, 89-108.
- Kettunen, K. et al. 2006. Analysis of EU Languages Through Text Compression [G] // I. Salakoski et al. *Advances in Natural Language Processing*. Berlin: Springer, 99-109.
- Kim, M., Crossley, S. A. & K. Kyle. 2017. Lexical Sophistication as a Multidimensional Phenomenon: Relations to Second Language Lexical Proficiency, Development, and Writing Quality [J]. *The Modern Language Journal* (1): 120-141.
- Kolmogorov, A. N. 1965. Three Approaches to the Quantitative Definition of Information [J]. *Problems of Information Transmission* (1): 1-7.
- Li, M. & P. M. Vitanyi. 1997. *An Introduction to Kolmogorov Complexity and Its Applications* [M]. New York: Springer.
- Li, M. et al. 2004. The Similarity Metric [J]. *IEEE Transactions on Information Theory* (12): 3250-3264.
- Linnarud, M. 1986. *Lexis in Composition: A Performance Analysis of Swedish Learners' Written English* [M]. Lund: CWK Gleerup.
- Liu, X., Li, F. & W. Xiao. 2024. Measuring Linguistic Complexity in Chinese: An Information-Theoretic Approach [J]. *Humanities and Social Sciences Communications* (1): 1-12.
- Lu, X. 2012. The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives [J]. *The Modern Language Journal* (2): 190-208.
- McWhorter, J. 2001. The World's Simplest Grammars Are Creole Grammars [J]. *Linguistic Typology* (2-3): 125-166.
- Miestamo, M. 2009. Implicational Hierarchies and Grammatical Complexity [G] // G. Sampson, D. Gil & P. Trudgill. *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press, 80-97.
- Montemurro, M. A. & D. H. Zanette. 2011. Universal Entropy of Word Ordering Across Linguistic Families [J]. *PLOS ONE* (6): e19875.
- Nichols, J. 1992. *Linguistic Diversity in Space and Time* [M]. Chicago: University of Chicago Press.
- Pallotti, G. 2015. A Simple View of Linguistic Complexity [J]. *Second Language Research* (1): 117-134.
- Paquot, M. 2019. The Phraseological Dimension in Interlanguage Complexity Research [J]. *Second Language Research* (1): 121-145.
- Sadeniemi, M. et al. 2008. Complexity of European Union Languages: A Comparative Approach [J]. *Journal of Quantitative Linguistics* (2): 185-211.
- Ströbel, M., Kerz, E. & D. Wiechmann. 2020. The Relationship Between First and Second Language Writing: Investigating the Effects of First Language Complexity on Second Language Complexity in Advanced Stages of Learning [J]. *Language Learning* (3): 732-767.

- Szmrecsanyi, B. 2015. Recontextualizing Language Complexity [G] // J. Daems et al. *Change of Paradigms-New Paradoxes: Recontextualizing Language and Linguistics*. Berlin: De Gruyter Mouton, 347-360.
- Ure, J. 1971. Lexical Density: A Computational Technique and Some Findings [G] // M. Coulard. *Talking about Text*. Birmingham: English Language Research, University of Birmingham, 27-48.
- Wang, G. et al. 2022a. Linguistic Complexity in Scientific Writing: A Large-scale Diachronic Study from 1821 to 1920 [J]. *Scientometrics* (128): 441-460.
- Wang, G. et al. 2022b. Kolmogorov Complexity Metrics in Assessing L2 Proficiency: An Information-Theoretic Approach [J]. *Frontiers in Psychology* (13): 1024147.
- Xiao, W. & S. Sun. 2020. Dynamic Lexical Features of PhD Theses Across Disciplines: A Text Mining Approach [J]. *Journal of Quantitative Linguistics* (2): 114-133.
- Ziv, J. & A. Lempel. 1977. A Universal Algorithm for Sequential Data Compression [J]. *IEEE Transactions on Information Theory* (3): 337-343.
- 鲍贵. 2008. 二语学习者作文词汇丰富性发展多纬度研究 [J]. 外语电化教学 (5): 38-44.
- 鲍贵. 2009. 英语学习者作文句法复杂性变化研究 [J]. 外语教学与研究 (4): 291-297+321.
- 陈蕊娜. 2017. 基于计量语言学指标的汉英文本特征比较研究 [D]. 杭州: 浙江大学.
- 胡初奋. 2021. 基于搭配的句法复杂度指标及其与汉语二语写作质量关系研究 [J]. 语言文字应用 (1): 132-144.
- 雷蕾. 2017. 中国英语学习者学术写作句法复杂度研究 [J]. 解放军外国语学院学报 (5): 1-10+159.
- 吴彤, 于金龙. 2003a. 柯尔莫哥洛夫: “复杂性研究中的欧几里德” [J]. 自然辩证法通讯 (6): 89-94.
- 吴彤, 于金龙. 2003b. 柯尔莫哥洛夫: 复杂性研究的逻辑建构过程评述 [J]. 自然辩证法研究 (9): 71-74.

Kolmogorov Complexity: Theories, Methods, Applications and Trends

XIAO Wei LIU Xun

Abstract: Kolmogorov complexity is an information theory-based language complexity measurement system that is able to measure the complexity at the morphological, syntactic and overall levels. It is characterized by its objectivity, economy, authenticity and comprehensiveness. Despite its potential, it remains largely unfamiliar to the domestic linguistic community. We begin by elucidating the theoretical underpinnings of Kolmogorov complexity and detailing the computational approaches to morphological, syntactical and overall complexity. Subsequently, we review the empirical applications of Kolmogorov complexity in linguistic research, noting an initial focus on language typology and a recent shift towards register variations, as well as a transition from ontology to second language studies. Ultimately, we anticipate future advancements in Kolmogorov complexity, particularly in enhancing semantic and pragmatic measurements, broadening its application scope, and intensifying the study of Chinese linguistics. Via this review we hope to deepen the understanding of Kolmogorov complexity among domestic scholars, enrich their measurement methods of language complexity, and promote the development of Chinese linguistics, Chinese-foreign language comparison and international Chinese language education research.

Key words: Kolmogorov complexity; linguistic complexity; measurement; information theory

责任编辑: 蒋勇军