基于熵算法的口译语言简化研究：受限语言视角*

香港理工大学　刘康龙　李德超

提要：本研究基于信息熵理论，采用词形熵和词性熵为核心指标分析英语口译语言、英语二语和英语母语在词汇和句法上的复杂度差异。研究结果表明：在词形熵方面，口译语言与二语无显著差别，相比母语均呈现出简化趋势；在词性熵方面，口译语言与二语同样无显著差异，且二者的词性熵均显著低于母语。本研究结果说明，口译语言和二语作为受限语言具有共性特征，在词汇选择和句法结构上均呈现出简化倾向。本研究通过引入“受限语言”这一视角，指出翻译过程与二语使用类似，都面临语言接触带来的双语激活、实时语言处理的认知负荷增加等挑战，因此二者在语言产出上呈现出相似特征，如简化和显化等。本研究不仅拓展了翻译语言作为“第三语码”的研究，更揭示了翻译语言的固有特征（如简化）并非单纯的翻译共性，而是与二语共享一定的认知加工机制，这些发现为探索翻译语言与其他接触语言形式之间的关联提供了理论和实践支持。

关键词：语料库、口译语言、受限语言、简化、信息熵

[中图分类号] H059　[文献标识码] A　[文章编号] 1000-0429（2025）01-0118-12

1. 引言

近年来，语音识别等技术的显著进步推动了基于语料库的口译研究的蓬勃发展。基于语料库的译文研究致力于探究译语的普遍特征，主要涉及简化、显化、规范化和匀质化等翻译现象（王克非、胡显耀 2008；庞双子、王克非 2018；蒋跃、马瑞敏、韩红建 2021）。口译研究亦循此范式，着重探讨口译语言的独特属性（Shlesinger 1998）。语料库分析能够高效处理大规模实际口译数据，因而

* 本文为李德超主持的香港研究资助局“优配研究金”（General Research Fund）资助项目“趋同还是趋异？基于语料库的学习者口译、专业口译和非母语语言种类的多维度分析”（15602621）以及“语料库驱动的限制性英语中的惯用语研究：多模态途径”（15603623）的阶段性成果。

在学界获得广泛关注（Shlesinger & Ordan 2012）。诸多研究表明，口译语言呈现出如简化（Xu & Li 2022；Xu & Liu 2023）、显化（Gumul 2021）以及受源语干扰（Ma & Cheung 2020）等特定倾向。此类研究主要对比源语与口译产出文本（Tang & Li 2016；符荣波、王克非 2021）、笔译文本与口译文本（Bernardini，Ferraresi & Miličević 2016；Xu & Li 2022），以及口译文本与自然口语文本（Liu，Cheung & Liu 2023；Xu & Liu 2023）。然而，尽管汉语语境下的口译研究已取得显著进展，但多局限于词汇和具体语法层面的探讨。这种研究视角的局限性导致研究结果不甚一致，其中关于简化现象的争议尤为突出。

翻译语言作为源语与目的语之间的“第三语码”（Frawley 1984），具有其独特的语言特征。系统探究这些特征可以更深入地理解翻译的本质。在此背景下，计量语言学以其严谨的方法论和对真实语料的系统分析（刘海涛 2018），为翻译研究开辟了新的研究路径，并逐渐受到关注。学者们通过句法复杂度、依存距离和活动度等量化指标系统揭示了译文的语言特征（徐佐浩、蒋跃 2021；吴继峰等 2023；范璐、蒋跃 2024）。计量分析方法也延伸至口译研究。研究者运用句法复杂度（Liu，Cheung & Liu 2023）、依存距离（Xu & Liu 2023，2024）和信息熵（Lin & Liang 2023）等指标，探究了口译语言的简化特征。这些研究为解析口译语言特性提供了新的理论视角和方法，也凸显了计量方法在口译研究中的应用价值。与此同时，受限语言（constrained language）的引入为口译研究提供了新的解释框架。该理论源自 Lanstyák & Heltai（2012）的论述，认为所有的语言产出都受到不同程度的限制，但某些语言类型会因特定交际语境而受到更显著的制约。Kruger & van Rooy（2016）进一步厘清了“受限语言”的概念，用以指代在特定交际语境中受到明显制约的语言产出。以翻译语言为例，它受到认知层面的双重制约：一方面来自双语激活机制，另一方面则受源语文本的限制。同时，翻译还需遵循目标语和目标文化的规范性要求。计量语言学方法与受限语言理论的结合，为深入探究口译语言特征提供了新的研究视角和方法论基础。

2. 口译简化研究述评

对口译简化现象的研究涵盖词汇、句法等多个语言层面，凸显了该语言现象的复杂性和多维特征。在 Laviosa（1998）的影响下，早期研究主要通过类符 / 形符比（TTR）和词汇密度等指标考察词汇层面的简化特征。Sandrelli & Bendazzoli（2005）基于欧洲议会口译语料库（European Parliament Interpreting Corpus，EPIC），研究了西班牙语和意大利语译入英语，以及西班牙语和英语译入意大利语四种语言组合。研究发现不同语言组合导致口译产出在词汇简化上存在显著差异，表明语言组合是影响口译词汇简化的重要因素。Russo，

Bendazzoli & Sandrelli（2006）发现，相比自然口语，西班牙口译语言的词汇密度更高，且高频词汇使用更为频繁。Xu & Li（2022）通过比较香港立法会的英语同传与笔译文本发现，口译文本在多项指标上呈现显著的简化特征。此外，与英国议会的英语母语演讲相比，口译文本的标准类符 / 形符比（STTR）和词汇多样性较低，但词汇密度却更高。这一看似矛盾的现象揭示了口译语言在词汇受限的情况下仍能维持较高的信息密度。在句法层面，Liu，Cheung & Liu（2023）通过分析 14 种句法复杂度参数，对比了同传英语、英语母语和英语二语的自然演讲，发现同传英语和二语相比母语均呈现出更为显著的简化特征。Xu & Liu（2023）基于依存语法框架，发现相比母语和二语的自然演讲，同传英语的平均依存距离更短，说明口译语言在句法结构上更为简化。此外，该研究还发现同传英语倾向于使用中心语后置的结构，该特征体现了汉语源语对译语词序的影响。Lv & Liang（2019）的研究发现，交传在信息密度、词汇重复率和复杂性上较同传更趋简化，表明其认知负荷并不低于同传。在跨语言比较研究方面，Bernardini，Ferraresi & Miličević（2016）通过对口译文本、笔译文本和原始演讲三种文本的两两对比，发现口译语言在词汇密度等维度上表现出更强的简化趋势。然而，不同的源语－目的语组合呈现出各自独特的简化特征：译入意大利语的口译语言倾向于降低词汇密度并缩短句长，译入英语的口译语言则主要表现为高频词覆盖率的提升。Dayter（2018）对俄英双向语料库的研究发现，俄语子库倾向于简化，而英语子库却呈现出相反特征，由此推测口译中语言组合是重要的影响因素。综上所述，口译中的简化特征受到口译类型、语言组合、口译方向等多种因素的影响，是一种动态且多维的语言现象，但目前仍缺乏全面和整体的评估指标，因此有必要借鉴其他学科开展深入研究。

3. 研究设计

3.1 研究问题

本研究采用信息熵指标，探究英语口译语言与英语母语和二语自然演讲在词汇和句法复杂度上的差异。研究主要围绕以下两个问题展开：1）基于信息熵分析，英语同声传译与英语母语和二语自然演讲在词汇和句法复杂度上是否存在显著差异？ 2）作为受限语言，英语同声传译与英语二语是否在某些方面具有相似特征？

3.2 语料库

本研究基于政治辩论英语可比语料库进行分析。该语料库由三个子语料库构成：汉译英的同传口译（IE）、英语二语的原始演讲（L2）和英语母语的原始演

讲（NE）。为保证研究的科学性和数据的可比性，我们选取了内容相近且可公开获取的英国议会和中国香港立法会辩论记录作为语料来源。英语二语和同传英语的讲话人均具有汉语粤方言背景，其中二语的数据来自两档电视访谈节目。在语料筛选过程中，我们确保主题须涵盖政治、社会和经济等多个领域，以降低主题差异对研究结果的影响，从而更准确地考察口译语言的简化特征。三个子语料库的数据均来自 2016—2020 年间，在语境、内容、规模和时间跨度等方面具有较高的可比性。每个子语料库均包含 65 篇经过词性标记的文本。

3.3 信息熵

信息熵是 Shannon（1948）提出的概念，用于量化和测算信息量。信息熵的基本原理是通过数学公式来衡量信息的随机性，该公式计算各种可能事件发生概率的对数与其概率乘积之和，详见公式（1）。在计算总熵时，事件的概率作为权重，其中高频事件对总熵的贡献较小，低频事件的贡献较大。公式（1）中，H 表示信息的总熵，P_i 代表特定事件发生的概率（通过相对频率计算），n 则是事件的总数。传统的类符 / 形符比仅衡量词汇多样性（即不同词汇的数量）；而信息熵不仅考虑词汇的出现频率，还关注词汇的分布特征。信息熵通过将词汇分布差异纳入考量，能够更精确地反映文本的复杂程度，从而提供更全面和准确的文本复杂度评估。

$$H = -\sum_{i=1}^{n} P_i \log_2 P_i \qquad (1)$$

信息熵自提出后在信息理论领域获得广泛应用（Shannon 1948，1951），并逐步延伸至语言学研究，被应用于文化多样性（Juola 2013）、作者识别（Khmelev 2000）及语言文本复杂度（Takahira，Tanaka-Ishii & Dębowski 2016）等议题的研究。在翻译研究领域，Liu，Liu & Lei（2022）运用信息熵分析发现，译文在词汇使用上趋向简化，而句法结构则呈现繁化特征。该研究表明熵值分析能有效揭示译文与原生文本的差异。信息熵作为一种定量分析工具，其优势在于能同时考虑词汇或词性的频次和分布特征，为文本复杂度测量提供更全面的视角。然而，该方法也存在局限性，如无法反映语义层面的复杂度，且对文本长度较为敏感。尽管信息熵在笔译文本分析研究中已有一定进展，但在口译研究领域仍处于起步阶段。本研究借鉴 Liu，Liu & Lei（同上）的方法，采用词形熵和词性熵来分别考察口译语言的词汇和句法复杂度。为提高分析精确度，我们使用基于 Python 的 SpaCy 工具进行词性标注，为每个词分配相应的语法类别。在自动标注基础上，我们进行了数据校对、错误修正和标记统一化处理。根据 Shi & Lei（2020）的建议，为确保熵值分析的准确性，我们将所有文本规范为 1500 词长，并去除标点符号和空格，以消除长度差异带来的影响。在此基础上，

我们计算了每个文本的词形熵和词性熵，通过统计分析探究三个子库在这两个指标上是否存在显著差异。

4. 分析结果

三个子库的词形熵和词性熵的统计结果显示，三个子库的词形熵均值相对接近，其中 NE 子库的词形熵均值略高于 IE 和 L2 子库，表明 NE 子库的词汇复杂度相对较高；在词性熵上，虽然三个子库的均值差异较小，但 NE 子库的均值略高，反映出 NE 子库的句法结构更为复杂。

图 1 和图 2 以箱线图的形式分别展示了三个子库在词形熵和词性熵上的分布特征。综合来看，NE 子库在词形熵和词性熵上均表现出更高的均值和较大的分布范围，而 IE 和 L2 子库则较为接近，分布更集中。

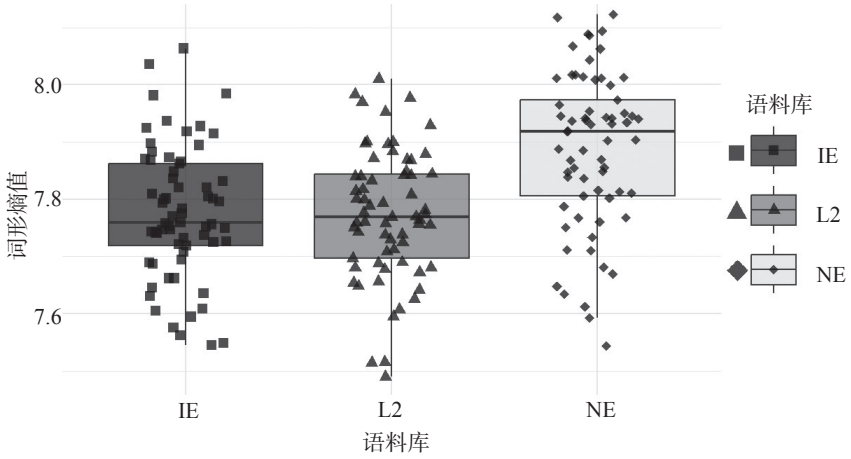

图 1. 三个子语料库的词形熵值分布

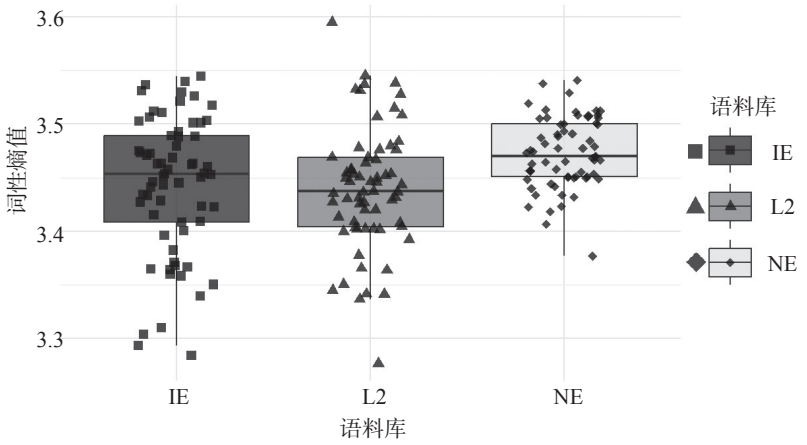

图 2. 三个子语料库的词性熵值分布

单因素方差分析显示，三个子库之间的词形熵值存在显著的组间差异 [$F(2, 192) = 17.94$，$p < 0.001$]，这说明至少一个子库的词形熵值与其他子库显著不

同。同样，词性熵值的方差分析也表现出统计显著性 [F（2，192）= 7.637，p < 0.001]。这两项检验结果均拒绝了零假设，即否定了三个子库在词形熵和词性熵均值上不存在差异的假设。这表明三个子库在词形和词性使用的复杂度方面存在显著差异。为进一步厘清子库间的差异模式，我们进行了事后多重比较分析（详见表 1）。Tukey 多重比较结果显示，L2 与 IE 的词形熵值无显著差异（p > 0.05），而 NE 与 IE（p < 0.001）、NE 与 L2（p < 0.001）之间存在极其显著的差异，这说明 NE 的词形熵值显著高于 IE 和 L2，而后两者之间无显著差异。词性熵值分析也呈现出相似结果：L2 与 IE 之间无显著差异（p > 0.05），但 NE 与 IE 之间差异显著（p < 0.05），而 NE 与 L2 之间的差异则更显著（p < 0.001），这进一步证实 NE 在词性使用上表现出更高的多样性和复杂性。

表 1. 三个子语料库词形熵和词性熵的 Tukey 多重比较结果

类别	比较	差异	下限	上限	校正后 p 值
词形熵	L2-IE	–0.003	–0.053	0.048	0.992
	NE-IE	0.110	0.059	0.161	< 0.001**
	NE-L2	0.113	0.062	0.163	< 0.001**
词性熵	L2-IE	–0.003	–0.026	–0.019	0.927
	NE-IE	0.030	0.008	0.052	< 0.005*
	NE-L2	0.033	0.011	0.055	< 0.001**

注： * 表示 p < 0.05，** 表示 p < 0.001。

5. 讨论

5.1 结果讨论

基于信息熵的分析表明，英语口译语言与英语母语和二语自然演讲在词汇和句法复杂度上均存在显著差异。在词汇复杂度上，口译语言与英语二语表现相似，二者均低于英语母语。词汇作为语言的基本单位，不仅是语言表达的基础要素，同时也是语法规则实现的载体。词汇熵值的计算为量化分析词汇多样性提供了一种客观方法（Chen，Liu & Altmann 2017）。较高的熵值反映词汇使用的多样性和丰富性，而较低的熵值则表明词汇重复度高，使用范围较窄。据此可推断，口译语言和二语较低的词形熵反映了词汇使用存在简化趋势。在口译语言中，这种现象主要源于口译过程中的时间压力和认知负载（Liu，Cheung & Liu 2023）。在同传任务中，口译员倾向于使用熟悉、常见且易于快速提取的词汇，这可能导致语言表达的复杂度降低。此外，由于中国香港特别行政区立法会的口译员均非英语母语者，其词汇选择可能呈现出一定的二语特征。本研究观

察到的二语词汇简化现象与香港的英语使用环境密切相关。虽然香港同时采用汉英双语作为官方语言,但正如 O'Brien(2004:1)所述:"粤英语码混搭(code mixing)一直是主导的语言使用模式。"这表明英语在香港并非独立或首选的交际媒介,而是更多地与粤语融合,形成混合语码。这种语言现象不仅存在于日常交谈,也广泛出现在媒体传播和教育领域。因此,英语在香港社会中实际更接近于外语角色,这一特点合理解释了本研究中二语词汇多样性偏低的现象。此外,口译语言和二语均属于受限语言范畴(Lanstyák & Heltai 2012),在跨语言和跨文化交际中,受限语言使用者为应对认知负荷并降低沟通风险,往往会有意或无意地简化语言表达(Kruger & van Rooy 2016;Liu, Cheung & Liu 2023)。

词性熵的分析结果显示,英语口译和二语在句法复杂度上表现相似,且均显著低于英语母语。词性作为词汇的语法属性分类,不仅反映词汇的语法功能,还体现词语间的句法关系。研究结果显示,口译语言的句法复杂度明显低于英语母语,这种现象可归因于同传的实时性和高认知负荷(Gumul 2021)。尽管口译员可能具备母语者水平的语言能力,但多任务处理的特性可能会影响其语言输出的句法复杂度(Gile 2009;Seeber 2011)。同时,由于需要同步监听和处理语言信息,认知负荷会相应增加,导致译员倾向于采用更简化的句法结构(Ma 2021)。英语二语的句法复杂度与口译语言相近,且显著低于英语母语。如前所述,中国香港地区的英语使用具有特殊性,其使用者倾向于采用简单句式和基础语法结构,以避免复杂语法可能带来的歧义或错误,从而确保交际清晰高效。此外,二语学习者的语言加工机制的自动化程度不如母语者,这也限制了其使用复杂句法的能力(Hinkel 2003;Liu, Cheung & Liu 2023)。

5.2 理论探讨

词形熵和词性熵的分析结果表明,英语同传和英语二语在词汇和句法层面都呈现出简化特征,这与母语者的英语形成鲜明对比。本研究发现与以往的 Liu, Cheung & Liu(2023)以及 Xu & Liu(2023)的研究结果相呼应。

正如前文所指出,口译语言和英语二语所表现的相似特征可从受限语言视角加以解释。传统翻译研究主要聚焦于语义转换的制约因素,往往以译文对原文语义的忠实度作为理论探讨的基础。而受限语言概念的引入则从理论层面揭示了翻译活动在认知层面上亦受限,并指出这种特性与其他接触语言(如二语)存在认知共性,从而导致不同类型的接触语言在表达或思维模式上呈现出相似性。"受限语言"概念的早期探讨可见于 House & Blum-Kulka(1986)的研究。她们通过比较翻译语言与二语写作的相似特征,推动了二语习得与翻译研究两个领域的交叉研究。Kruger & van Rooy(2016:27)进一步阐明,受限

语言是在具有明显制约条件的交际环境中形成的语言形式。因此，研究者常将译语与其他非母语或经介导（mediation）的语言类型（如学习者语言、经编辑语言、新英语变体等）进行比较，以探究其共性（Kruger 2012；Kruger & De Sutter 2018）。该概念有效整合了翻译研究、接触语言学、二语习得和双语研究。近年来，基于受限语言框架的实证研究从多个维度证实，各种形式的受限语言因面临相似的制约而表现出共性特征，例如口译语言和笔译语言在词汇使用、句法结构和篇章衔接等方面与二语表达高度相似（Kajzer-Wietrzny & Ivaska 2020；Ivaska，Ferraresi & Bernardini 2022；Kajzer-Wietrzny 2022）。在口译语言和二语环境中，由于认知负荷加重、目标语言掌握程度不足以及实时处理压力等因素，语言使用者倾向于简化语言结构（De Clercq & Housen 2017）。这种简化既体现在选用常见的简短词语，也表现为构建较为简洁的句法结构。本研究在一定程度上支持这一论断。

本研究观察到的口译语言在词汇和句法层面的简化现象，可借助 Gile（2009：160—165）的口译认知负荷模型（Effort Models of Interpreting）进行解释。该理论模型指出，同传过程中口译员需要同时处理多重任务，包括听辨、记忆、产出和协调，这些任务的总和构成了口译的认知负荷。在同传的高压环境下，口译员需同步处理输入信息，并准确、快速地输出目标语言。为有效管理有限的认知资源并降低认知负荷，口译员可能会不自觉地简化词汇和句法结构，以减轻语言产出和短期记忆的负荷，从而为听辨和任务协调腾出更多认知资源。简化的语言特征因需较少认知处理，有助于保持口译的流畅度和更准确地传递信息。研究表明，对于双语者而言，即使主要使用一种语言进行交际，另一种语言也会在认知层面保持活跃状态（Shook & Marian 2012）。在这种双语激活的状态下，口译员需要启动复杂的认知控制机制，不仅涉及语言知识的检索和运用，还包括对非目标语言的抑制以及对两种语言激活水平的调节。本研究的结果表明，口译语言的简化特征在很大程度上受到这种双语认知加工过程所带来的制约。

6. 结语

本研究借助可比语料库范式，采用信息熵的方法，深入对比分析了受限语言与非受限语言在词汇和句法复杂度方面的差异。研究结果表明，受限语言的词汇和句法复杂度较低，有力地支持了简化的理论假设。此外，研究还发现，两种受限语言在词汇和句法复杂度上有较高相似性。这些发现不仅拓展了对翻译作为“第三语码”的研究范畴，还通过探索翻译语言与其他类型接触语言的共性特征，表明翻译特征并非仅由语言介导造成，而是源自多重因素的共同作用（Kotze & van Rooy 2024）。针对这些发现，本研究就口译和二语处理中的认知

负荷及源语言干扰效应等问题进行了理论阐释。

本研究也存在若干局限。首先，使用信息熵评估词汇和句法复杂度时主要依赖量化特征，未能质性地考察句法结构和深层语义的复杂度。未来研究可结合其他指标，更准确地衡量文本的风格及复杂度。其次，研究语料主要来自中国香港地区，这可能会影响结果的普适性。为克服上述局限，未来研究可从数个维度拓展：首先，可纳入来自其他地区或更广泛语言背景的二语及口译语料，以提高数据的多样性和代表性；其次，可引入更多计量语言学指标，深入探讨受限语言与非受限语言在复杂度方面的异同；最后，可对源语言变量及其“透射”或干扰效应（Teich 2003）进行细致分析，以揭示源语对目标语生成的影响。这种跨学科研究方法将推动翻译学发展，促进翻译学、语言学和认知科学的交叉融合，为理解翻译这一独特的语言处理方式提供新视角。

参考文献

Bernardini, S., A. Ferraresi & M. Miličević. 2016. From EPIC to EPTIC — Exploring simplification in interpreting and translation from an intermodal perspective [J]. *Target* 28(1): 61–86.

Chen, Ruina, Haitao Liu & G. Altmann. 2017. Entropy in different text types [J]. *Digital Scholarship in the Humanities* 32(3): 528–542.

Dayter, D. 2018. Describing lexical patterns in simultaneously interpreted discourse in a parallel aligned corpus of Russian-English interpreting (SIREN) [J]. *Forum* 16(2): 241–264.

De Clercq, B. & A. Housen. 2017. A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity [J]. *The Modern Language Journal* 101(2): 315–334.

Fan, Lu & Yue Jiang [范璐、蒋跃]. 2024. “In-between Hypothesis” — A novel proposition on the feature of translational language: A quantitative study based on dependency grammar [J]. *Foreign Language Teaching and Research* (1): 136–147. [翻译语言特征新假设“折中假设”——基于依存语法的计量研究，《外语教学与研究》1]

Frawley, W. 1984. Prolegomenon to a theory of translation [A]. In W. Frawley (ed.). *Translation: Literary, Linguistic and Philosophical Perspectives* [C]. London: Associated University Press. 250–263.

Fu, Rongbo & Kefei Wang [符荣波、王克非]. 2021. Lexical patterns in interpreted and spontaneous English speeches: A comparable, intermodal and corpus-based study [J]. *Foreign Language Teaching and Research* (6): 912–923. [基于跨模式类比语料库的汉英口译词汇特征研究，《外语教学与研究》6]

Gile, D. 2009. *Basic Concepts and Models for Interpreter and Translator Training* [M]. Amsterdam: John Benjamins.

Gumul, E. 2021. Explicitation and cognitive load in simultaneous interpreting: Product- and process-oriented analysis of trainee interpreters’ outputs [J]. *Interpreting: International Journal of Research and Practice in Interpreting* 23(1): 45–75.

Hinkel, E. 2003. Simplicity without elegance: Features of sentences in L1 and L2 academic texts [J].

TESOL Quarterly 37(2): 275–301.

House, J. & S. Blum-Kulka. (eds.). 1986. *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies* [C]. Tübingen: G. Narr.

Ivaska, I., A. Ferraresi & S. Bernardini. 2022. Syntactic properties of constrained English: A corpus-driven approach [A]. In S. Granger & M. Lefer (eds.). *Extending the Scope of Corpus-based Translation Studies* [C]. London: Bloomsbury Academic. 133–157.

Jiang, Yue, Ruimin Ma & Hongjian Han [蒋跃、马瑞敏、韩红建]. 2021. A quantitative analysis of "the third code" at the syntactic level [J]. *Foreign Language Teaching and Research* (6): 830–841. [句法层面"第三语码"的计量研究，《外语教学与研究》6]

Juola, P. 2013. Using the Google N-gram corpus to measure cultural complexity [J]. *Literary and Linguistic Computing* 28(4): 668–675.

Kajzer-Wietrzny, M. 2022. An intermodal approach to cohesion in constrained and unconstrained language [J]. *Target* 34(1): 130–162.

Kajzer-Wietrzny, M. & I. Ivaska. 2020. A multivariate approach to lexical diversity in constrained language [J]. *Across Languages and Cultures* 21(2): 169–194.

Khmelev, D. 2000. Disputed authorship resolution through using relative empirical entropy for Markov chains of letters in human language texts [J]. *Journal of Quantitative Linguistics* 7(3): 201–207.

Kotze, H. & B. van Rooy. 2024. Introduction: The constrained communication framework for studying contact-influenced varieties [A]. In B. van Rooy & H. Kotze (eds.). *Constraints on Language Variation and Change in Complex Multilingual Contact Settings* [C]. Amsterdam: John Benjamins. 1–28.

Kruger, H. 2012. A corpus-based study of the mediation effect in translated and edited language [J]. *Target* 24(2): 355–388.

Kruger, H. & G. De Sutter. 2018. Alternations in contact and non-contact varieties: Reconceptualising *that*-omission in translated and non-translated English using the MuPDAR approach [J]. *Translation, Cognition & Behavior* 1(2): 251–290.

Kruger, H. & B. van Rooy. 2016. Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English [J]. *English World-Wide* 37(1): 26–57.

Lanstyák, I. & P. Heltai. 2012. Universals in language contact and translation [J]. *Across Languages and Cultures* 13(1): 99–121.

Laviosa, S. 1998. Core patterns of lexical use in a comparable corpus of English narrative prose [J]. *Meta* 43(4): 557–570.

Lin, Yumeng & Junying Liang. 2023. Informativeness across interpreting types: Implications for language shifts under cognitive load [J]. *Entropy* 25(2): Article No. 243.

Liu, Haitao [刘海涛]. 2018. *Advances in Quantitative Linguistics* [M]. Hangzhou: Zhejiang University Press. [《计量语言学研究进展》。杭州：浙江大学出版社]

Liu, Kanglong, Zhongzhu Liu & Lei Lei. 2022. Simplification in translated Chinese: An entropy-based approach [J]. *Lingua* 275: Article No. 103364.

Liu, Yi, Andrew K. F. Cheung & Kanglong Liu. 2023. Syntactic complexity of interpreted, L2 and L1 speech: A constrained language perspective [J]. *Lingua* 286: Article No. 103509.

Lv, Qianxi & Junying Liang. 2019. Is consecutive interpreting easier than simultaneous interpreting? — A corpus-based study of lexical simplification in interpretation [J]. *Perspectives* 27(1): 91–106.

Ma, Xingcheng. 2021. Coping with syntactic complexity in English-Chinese sight translation by translation and interpreting students. An eye-tracking investigation [J]. *Across Languages and Cultures* 22(2): 192–213.

Ma, Xingcheng & Andrew K. F. Cheung. 2020. Language interference in English-Chinese simultaneous interpreting with and without text [J]. *Babel* 66(3): 434–456.

O'Brien, T. 2004. Writing in a foreign language: Teaching and learning [J]. *Language Teaching* 37(1): 1–28.

Pang, Shuangzi & Kefei Wang [庞双子、王克非]. 2018. The explicitness of register features in literary translations: A diachronic study [J]. *Chinese Translators Journal* (5): 13–20, 48. [翻译文本语体"显化"特征的历时考察,《中国翻译》5]

Russo, M., C. Bendazzoli & A. Sandrelli. 2006. Looking for lexical patterns in a trilingual corpus of source and interpreted speeches: Extended analysis of EPIC (European Parliament Interpreting Corpus) [J]. *Forum* 4(1): 221–254.

Sandrelli, A. & C. Bendazzoli. 2005. Lexical patterns in simultaneous interpreting: A preliminary investigation of EPIC (European Parliament Interpreting Corpus) [A]. In P. Danielsson & M. Wagenmakers (eds.). *Proceedings from the Corpus Linguistics Conference Series* (Vol. 1) [C]. Birmingham: University of Birmingham. 1–24.

Seeber, K. 2011. Cognitive load in simultaneous interpreting: Existing theories — New models [J]. *Interpreting* 13(2): 176–204.

Shannon, C. 1948. A mathematical theory of communication [J]. *Bell System Technical Journal* 27(3): 379–423.

Shannon, C. 1951. Prediction and entropy of printed English [J]. *Bell System Technical Journal* 30(1): 50–64.

Shi, Yaqian & Lei Lei. 2020. Lexical richness and text length: An entropy-based perspective [J]. *Journal of Quantitative Linguistics* 29(1): 62–79.

Shlesinger, M. 1998. Corpus-based interpreting studies as an offshoot of corpus-based translation studies [J]. *Meta* 43(4): 486–493.

Shlesinger, M. & N. Ordan. 2012. More *spoken* or more *translated*? Exploring a known unknown of simultaneous interpreting [J]. *Target* 24(1): 43–60.

Shook, A. & V. Marian. 2012. Bimodal bilinguals co-activate both languages during spoken comprehension [J]. *Cognition* 124(3): 314–324.

Takahira, R., K. Tanaka-Ishii & Ł. Dębowski. 2016. Entropy rate estimates for natural language — A new extrapolation of compressed large-scale corpora [J]. *Entropy* 18(10): Article No. 364.

Tang, Fang & Dechao Li. 2016. Explicitation patterns in English-Chinese consecutive interpreting: Differences between professional and trainee interpreters [J]. *Perspectives* 24(2): 235–255.

Teich, E. 2003. *Cross-linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts* [M]. Berlin: De Gruyter Mouton.

Wang, Kefei & Xianyao Hu [王克非、胡显耀]. 2008. A parallel corpus-based study on lexical features of translated Chinese [J]. *Chinese Translators Journal* (6): 16–21. [基于语料库的翻

译汉语词汇特征研究，《中国翻译》6]

Wu, Jifeng, *et al*. [吴继峰等]. 2023. A comparative study of the syntactic complexity of translated Chinese and original Chinese [J]. *Foreign Language Teaching and Research* (2): 264–275. [翻译汉语和原创汉语句法复杂度对比研究，《外语教学与研究》2]

Xu, Cui & Dechao Li. 2022. Exploring genre variation and simplification in interpreted language from comparable and intermodal perspectives [J]. *Babel* 68(5): 742–770.

Xu, Han & Kanglong Liu. 2023. Syntactic simplification in interpreted English: Dependency distance and direction measures [J]. *Lingua* 294: Article No. 103607.

Xu, Han & Kanglong Liu. 2024. The impact of directionality on interpreters' syntactic processing: Insights from syntactic dependency relation measures [J]. *Lingua* 308: Article No. 103778.

Xu, Zuohao & Yue Jiang [徐佐浩、蒋跃]. 2021. Activity of translational Chinese: A study based on three online corpora [J]. *Foreign Language Teaching and Research* (1): 113–123. [翻译汉语的活动度——基于在线语料库的研究，《外语教学与研究》1]

An Entropy-based Study of Simplification in Interpreting: A Constrained Language Perspective

LIU Kanglong LI Dechao

(Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong 999077, China)

Abstract: Drawing on information entropy theory and using wordform entropy and part-of-speech entropy as key indicators, this study analyzes the differences in lexical and syntactic complexity between interpreted English, English as a second language, and native English. In terms of wordform entropy, the findings reveal that interpreted language does not differ significantly from second language output, with both exhibiting simplification tendencies compared to native language. Similarly, for part-of-speech entropy, interpreted and second language exhibit no significant differences, and both are significantly lower than native language. These findings suggest that as constrained languages, interpreted and second languages share common traits and exhibit simplification in lexical choice and syntax. By introducing the concept of "constrained language," this study argues that both translation processes and second language use face similar challenges: bilingual activation during language contact and heightened cognitive load from real-time processing, which together result in shared linguistic features such as simplification and explicitation. This research not only expands the study of translated language as a "third code" but also reveals that the inherent features of translated language (e.g., simplification) are not merely translation universals but rather reflect cognitive processing mechanisms shared with second language production. These findings provide both theoretical and empirical support for understanding the relationship between translated language and other forms of contact language.

Keywords: corpus, interpreted language, constrained language, simplification, information entropy

收稿日期：2024-08-29；修改稿 2024-11-13；本刊修订 2024-11-26

通讯地址：999077 香港 香港理工大学中文及双语学系

电子邮箱：kl.liu@polyu.edu.hk

View publication stats