《童年》汉译语言特征计量分析

关秀娟 那峻源

（黑龙江大学俄语学院，哈尔滨 150080）

提　要：本研究以数字人文为切入点，采用语料库研究范式对《童年》的汉译本进行语言特征对比研究，为文学作品的汉译研究新路径探索提供参考。通过收集原文本和汉译本语料建立语料库，分析俄汉语料的宏观和微观特征，进而探究汉译本语言特征的异同。数据结果表明，汉译本在连词和动词搭配范式的计量特征方面呈现高度相似性，表明其语言依存结构上差距较小；在词汇丰富度、可读性和主题性检验上呈现一定的差异性，表明汉译本的语言复杂度、可读性和主题性表达差距较大。汉译本相较于原著，其连词翻译总体呈现隐化倾向，在动词计量特征方面，汉译本的动词—介词短语搭配呈现出隐化的倾向，动词拷贝结构数量显著大于原著，汉译本更注重于动词本身的意义传达，而原著更注重与其他词的搭配来进行意义传达。

关键词：《童年》；汉译；语言特征；计量分析；对比

中图分类号：H059　　文献标识码：A

交叉学科的发展推动了国外文学汉译语言特征的研究，尤其是利用语料库等技术工具的研究。然而，目前基于语料库的文学汉译语言特征研究还存在一些不足。其中，计量分析的数据类型较为有限，而且大多数研究集中于对宏观或微观语言特征的单一维度分析。很少有研究同时对宏观和微观语言特征展开综合计量分析。另外，在计量统计方面也有改进的空间，如对词汇丰富度的研究主要聚焦于类符/形符比和标准类符/形符比的计算上，未能引入更为复杂精准的词汇丰富度计算公式，在统计方面计量维度也较为单一等。本文以高尔基作品《童年》的原著以及杨实和郑海凌两个汉译本[1]为研究对象。研究中使用了 python 以及自然语言处理工具 spaCy[2] 将《童年》的原著以及两个汉译本建成语料库，并对语料库同时进行宏观和微观语言特征计量分析。宏观语言特征研究涵盖多个计量维度，包括词汇丰富度、文本可读性和主题性，同时引入了多种计算公式，以确保数据的可靠性。在微观语言特征方面，本研究对具体词汇的语言特征进行了分析，并提出了十余种词汇计量特征，以确保研究的广度。由于篇幅限制，本研究仅以连词、动词为核心词进行计量分析。在未来研究中，可以进一步扩展分析的维度，以获得更全面的研究结果。

1 宏观语言特征计量分析

宏观语言特征是指不涉及具体细微语言的计量特征，关注文本的总体结构、语言风格，主要包括词汇丰富度、可读性和主题性。

1.1 词汇丰富度计量分析

词汇丰富度（lexical diversity）是语料的宏观特征之一，能够表现出语料语言多样性、信息丰富度和译者风格等。本研究计算了形符数、类符数、类符/形符比（TTR）、标准类符

/形符比（STTR）、杜加斯特指数（Dugast's Uber Index）和赫丹 Vm (Herdan Vm）指数。标准类符/形符比（STTR）是基于 TTR 公式的改良版本，即通过计算每 n 个字符的 TTR，最后再计算所有 TTR 的均值，以减少误差，其公式为：

$$STTR = \sum_{i=0}^{n} TTRi/n$$

表 1 词汇丰富度计量特征表

	原著	杨译本	郑译本
形符数	54,567	51,980	81,453
类符数	7,665	8,209	11,135
类符/形符比	14.05%	15.79%	13.67%
标准类符/形符比	42.05%	39.09%	49.99%
杜加斯特指数	60.61	63.88	64.25
赫丹 **Vm** 指数	0.15	0.124	0.147

经计算得到 7 组计量特征数据用于描述语料的词汇丰富度。观察表 1 数据可知，杨译本形符数仅为原著的 95.25%，而一般文学作品的非删减汉译本形符数都要高于原著，因此可以判断杨译本是删减版本。郑译本则较为忠实地保留了原著的词汇丰富度特征，甚至优于原著。汉译本的类符数均大于原著，这说明汉译本翻译有显化趋势。接下来使用皮尔逊相关系数对比了汉译本的标准类符/形符比、杜加斯特指数和赫丹 Vm 指数的相关性，随后进行了相关性检验。结果显示，其皮尔逊相关系数高达 0.999，$p<0.001$，显示出此三个计量特征具有强正相关性。这表明郑译本使用了更丰富的词汇，在不考虑读者阅读能力的情况下，郑译本能够提供更充实、信息量更密集的阅读体验。

1.2 文本可读性计量分析

文本可读性（readability），即文本的阅读难易程度，是一种重要的计量特征。其中，平均句长（mean sentence length）是一个基础且关键的指标，用于在句法和语篇层面进行比较分析，以评估阅读体验。在本研究中，使用 spaCy 对原著和汉译本的平均句长和平均依存距离进行计量分析。与 TTR 计算方式类似，由于在大文本情境下直接计算平均句长具有不可避免的误差，因此本研究改进了计算方法，采用下列标准平均句长（standardized mean sentence length）的公式：

$$SMSL = \sum_{i=0}^{n} MSLi/n$$

在这个公式中，每隔 n 个字符计算一次平均句长，然后取这些值的算术平均数。本研究中，选择 n 值为 1000 来计算 SMSL 值。此外，还对语料的依存数量和平均依存距离进行了计算，6 个计量特征数据如下表所示：

表 2 平均句子长度与依存距离数据

	原著	杨译本	郑译本
形符数	54,567	51,980	81,453
句子总数	2,587	4,483	6,555
平均句长	21.09	11.59	12.42
标准平均句长	14.82	10.24	10.94
依存数量	58,729	56,295	84,542
平均依存距离	3.42	3.49	3.29

根据上表数据分析，可以得出以下结论：在平均句长这一计量特征中，杨译本<原著<郑译本。在句子总数这一计量特征中，原著为 2,587 个，杨译本为 4,483 个，而郑译本的句子总数最多，为 6,555 个。这两项数据表明，汉译本在对原著进行翻译时采用了较多的句子分割的处理方式，即将一个长句处理为两个及以上的短句的翻译方式，且郑译本则更加频繁地使用了这一翻译方法。在平均句长与标准平均句长两项数据方面，郑译本与杨译本并无太大差别，说明郑译本在句子层面可读性与杨译本保持一致。但郑译本的句子总数远高于杨译本，郑译本在语篇的可读性方面的阅读难度要大于杨译本。

此外，从平均依存距离计量特征来看，郑译本<原著<杨译本，方差为 0.0089。这表明三个语料的平均依存距离较为集中一致，差异较小。结合标准平均句长这一数据，可以进一步得出以下结论：原著的平均句子长度较长，但在保持相似的依存距离的前提下，原著中的每个句子包含更多的依存关系和搭配信息。而汉译本的平均句子长度和平均依存距离较为接近，表明在这两个计量特征上汉译本呈现出相对一致的特征。然而，郑译本的依存数量远远超过了杨译本，这说明在依存关系的层面上，郑译本的文本具有更高的阅读难度，对于读者而言更有挑战性。

在可读性的计量方面对原著采用了三个较为权威的可读性公式，分别为：Dale-chall、Flesch-Kincaid Grade Level 和 Flesch Reading Ease，这三个公式基于高频词、高难词、复合句等数据进行计算，适用于普通文学文本的可读性计算。对于汉译本使用了第三方库 cntext 的可读性计算公式进行计算，计算过程中将三个语料平均分为 5 块后取平均值，其可读性结果如下所示：

表 3 原著可读性指标（均值）

	Dale-Chall	**Flesch-Kincaid**	**Flesch Reading Ease**
得分	8.79	13.92	67.38
对应国外年级	11-12 年级	11-18 年级	8-9 年级

表 4 汉译本可读性指标（均值）

	扬译本	郑译本	T 值	P 值	显著性
可读性 1	24.05	28.74	-7.366	<0.001	√
可读性 2	0.071	0.068	1.34	0.21	×
可读性 3	11.92	14.40	-7.340	<0.001	√
对应国内年级	初中 1—2 年级	初中 3 年级—高中 2 年级	—	—	—

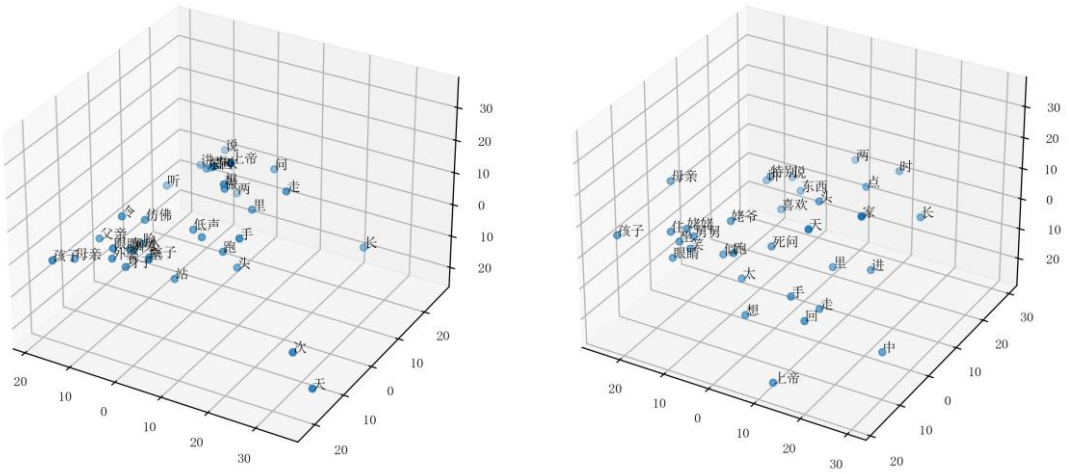
通过表 3 与表 4 的可读性计量特征可知：原著的阅读难度相当于母语者 11 年级水平，而杨译本的难度相当于初中 1—2 年级，郑译本则介于初中 3 年级—高中 2 年级之间。在对汉译本的可读性样本数据进行独立样本 T 检验和显著性检验后发现，除了可读性 2 之外，可读性 1 与可读性 3 这两个计量特征具有统计学上的显著性（$p<0.001$）。这表明杨译本的阅读难度明显低于郑译本。

综上所述，从句法计量特征分析来看，通过考察平均句长、平均依存距离等数据，发现汉译本可读性具有高度相似性，说明从句子层面并不能看出二者的阅读难度差别。通过可读性计算公式数据可知，汉译本在语篇层面上体现出了显著差异。杨译本整体可读性对应的年级范围为国内初中 1—2 年级，而郑译本稍高，对应初中 3 年级—高中 2 年级。这些指标说明在整体语篇的角度来看，杨译本的阅读难度较低，可读性高于郑译本。郑译本相对而言存在更多的高难表达，这对读者的阅读水平要求更高。

1.3 主题性计量分析

主题性是语料的宏观特征之一，词频（Term Frequency）是其重要计量特征之一。本研

究忽略不具备实际意义的虚词，在统计时仅统计有实际意义的词汇的词频，通过对比三个语料的高频词，使用 spaCy 计算三个语料中高频词的词向量，利用 PCA 主成分分析将词向量降低至三维，再用 matplotlib 将其可视化，由于篇幅限制，只取部分高频词进行展示，结果如下图所示：



(a) **(b)**

图 1 汉译本主题词 PCA 分析三维对比图

在杨译本中（图 1a），“姥姥”“姥爷”和“母亲”这类词语表明了小说中关键角色的称谓，它们在词向量空间中紧密聚集，形成一个称谓语聚类。这在词向量空间中构建了一个语义场景，反映了小说的相关主题性，即这些称谓语聚类中的角色在主人公阿廖沙的童年生活中扮演着重要的角色。在郑译本中，主成分分析（图 1b）也展现出了类似的上述称谓词的聚类，表明这两种译本在表达这一特定主题方面具有一定的相似性。

此外，动词如“说”“想”“讲”“笑”“死”和“做”构成了一个动词的聚类，称谓词是这些动词的主要依存对象。这个聚类包含较少与情感相关的表达，主要由描述事实的词汇组成，在一定程度上反映了小说的整体基调。从图 1 可以观察到，在称谓词和动词的聚类方面，郑译本在词向量空间内相对于杨译本更为紧密。除了上述聚类，还存在一些独立于聚类之外的高频词。如郑译本中，高频词汇“上帝”与其他词汇的距离更近，表明它与其他主题词汇的联系更为紧密，而在杨译本中“上帝”与词聚类的联系较为松散，表明这些具备特殊意向的词汇方面，郑译本更加贴合原著。

综合而言，汉译本在主题高频词的词类方面差异较小，但郑译本在主题高频词的词向量空间分布方面表现更佳，情感表达较为贴合原著，且主题词汇之间的联系更为紧密。在词汇聚类方面，郑译本的主题性表达更为精确和连贯，有助于读者更好地理解人物和背景描述，相较之下，杨译本则稍显不足。

2 微观语言特征计量分析

微观语言特征是指语料中具体的元素或特征，包括词汇特征、语法特征、修辞特征等。由于篇幅限制，微观语言特征中仅以连词和动词为核心词考察其计量特征，计算了连词和动词的核心依存关系的搭配范式以及频次，以期寻找汉译本的语言风格的异同。

2.1 连词特征计量分析

连词扮演着重要的逻辑连接角色。因此，连词的处理方式对于翻译的流畅性至关重要。本研究对原著和汉译本的连词进行计量分析，结果表明，原著共 26 种连词，两个汉译本均

为 36 种。就连词的丰富度而言，汉译本要高于原著。从连词数量角度来看，原著显著多于汉译本，如原著中仅连词“и”就出现了 1953 次，甚至高于汉译本连词数量的总和。具体的连词搭配情况详见下表：

表 5 连词搭配计量特征

原著		杨译本		郑译本	
搭配范式	数量（占比）	搭配范式	数量（占比）	搭配范式	数量（占比）
连词+动词	1,086 (55.12%)	连词+名词	141 (54.23%)	连词+名词	245 (54.32%)
连词+名词	465 (23.6%)	连词+动词	96 (36.9%)	连词+动词	168 (37.25%)
连词+形容词	196 (9.94%)	连词+专有名词	14 (5.38%)	连词+专有名词	27 (5.98%)
连词+副词	123 (6.24%)	连词+代词	5 (1.92%)	连词+协同连词	1 (0.22%)
连词+代词	27 (1.37%)	连词+协同连词	1 (0.38%)	连词+代词	3 (0.66%)
连词+专有名词	42 (2.13%)	连词+形容词	2 (0.76%)	连词+数词	2 (0.44%)
连词+协同连词	4 (0.2%)	连词+数词	1 (0.38%)	连词+副词	2 (0.44%)
连词+特殊词	8 (0.4%)	—	—	连词+形容词	2 (0.44%)
连词+限定词	5 (0.25%)	—	—	连词+介词	1 (0.22%)
连词+助动词	2 (0.1%)	—	—	—	—
连词+数词	8 (0.4%)	—	—	—	—
连词+从属连词	2 (0.1%)	—	—	—	—
连词+感叹词	1 (0.05%)	—	—	—	—
连词+介词	1 (0.05%)	—	—	—	—
总计	1,970		260		451

通过表 5 数据可知，汉译本的连词搭配范式相比原著有较大差异。首先就搭配范式的种类而言，原著更为丰富，共 14 种连词搭配，其中具有统计学意义的搭配共 6 种：连词+动词、连词+名词、连词+形容词、连词+副词、连词+代词和连词+专有名词，共占所有搭配范式的 98.4%，其余搭配因数量太少，无法确定其结果或差异是否是由随机偶然事件引起，因此不作探讨。郑译本共包含 9 种连词搭配，杨译本共 7 种，二者具有统计学意义的搭配均共 3 种，分别是连词+名词、连词+动词和连词+专有名词，分别占所有搭配范式的 96.51%和 97.8%。在搭配范式数量方面，两个汉译本分别只有 260 个和 451 个连词搭配范式，明显低于原著的 1970 个。这表明汉译本倾向于简化这些连接词搭配，以减少语料信息密度和难度，从而增加信息传播的广度。原著中，连词+动词的搭配数量最多，而汉译本中连词+名词的搭配范式数量最多。这表明在连接词的依存对象方面，原著与汉译本的语言特征存在显著差异：原著更强调动词的“动态”，而汉译本则更倾向于名词的“静态”。原著中连词后的形容词和副词在汉译本中的数量显著下降，这可能是因为汉语的定语修饰词位置较为灵活，可以脱离连词的约束范围而存在导致的。此外，汉译本在连词+代词、连词+协同连词、连词+形容词、连词+介词的搭配范式数量也显著低于原著，说明汉译本在这些搭配范式的翻译中呈现隐化的特征。

2.2 动词特征计量分析

本研究针对三个语料中的动词进行了系统研究，在计量方面，重点对其依存关系进行了计算与统计，核心依存关系包括：动词—形容词性修饰词—宾语（记为 v-amod-obj）、动词—名词主语（记为 v-nsubj）、动词—介词短语（记为 v-adp-n）、动词—副词（记为 v-advmod）、动词—动词（记为 v-v）、被动动词（记为 v-pass），共 6 项数据。使用 spaCy 对三个语料进行分析后，其结果如下表所示：

表 6 动词搭配计量特征

搭配范式	原著	杨译本	郑译本
v-amod-obj	146 (3.15%)	48 (0.82%)	95 (0.96%)
v-nsubj	2,094 (45.24%)	2,765 (47.62%)	4,463 (45.21%)
v-adp-n	1,122 (24.24%)	266 (4.58%)	516 (5.22%)
v-advmod	593 (12.8%)	486 (8.37%)	905 (9.16%)
v-v	452 (9.76%)	2,154 (37.09%)	3,711 (37.59%)
v-pass	221 (4.77%)	87 (1.71%)	181 (1.83%)
总计	4,628	5,806	9,871

首先，对 6 项动词搭配范式数量进行分析。可以看到，郑译本的动词搭配数量是原著的约 2.1 倍，而杨译本仅为原著的约 1.25 倍。这表明郑译本在表示主体动作行为表达时有显著的显化特征，而杨译本则无明显显化特征。原著中，动词—名词主语和动词—介词短语搭配范式分别占 45.24%和 24.24%，这两种搭配在原著中具有最高的频次，是最为重要且频繁出现的两类搭配。两个汉译本中的动词—名词主语搭配也占据了较大的比例，其频次和占比均超过了原著，表明汉译本均有将原著的人称代词显化的特征。汉译本的动词—介词短语搭配数量显著低于原著，说明汉译本将原著的介词结构进行了隐化处理。在动词—动词（连续动词结构或动词拷贝结构）搭配数量方面，原著与两个汉译本之间存在显著差异。两个汉译本中连续动词结构的数量占比分别为 37.09%和 37.59%，远高于原著的 9.76%。通过统计对比分析，发现 v-adp-n 与 v-v 这两组数据之间存在较大差距，皮尔逊相关系数为-0.9993，p 值为 0.0238。这表明这两组样本数据在统计学上具有显著性差异（$p<0.05$），即汉译本中动词拷贝结构越多，动词—介词短语搭配的数量相应越少，这说明汉译本在介词的使用方面不如原著灵活，更加突出动词本身的意义传达，而原著的部分动词则需要借助介词才能完成完整的意义表达。此外，原著形容词和副词的运用与动词更为紧密，这是由形容词、副词在两种语言中的差异导致的，汉语形容词和副词可以直接作谓语或谓语核心，动词成分要为形容词、副词让步，间接导致了汉译本中 v-amod-obj 和 v-advmod 搭配范式数量的下降。

综上所述，在动词使用方面原著与汉译本呈现出显化与隐化的不同特征。郑译本的主体动作行为表达更为突出，杨译本与原著较为接近。汉译本均有大量动词拷贝结构，不需借助过多介词来进行意义表达，因此动词—介词短语搭配显著少于原著。此外，汉译本形容词和副词与动词的联系不如原著紧密，在表达此类修饰意义时有显著差异。在原著中存在较多被动动词，而在汉译本中，动词主要以主动态呈现。

3 结语

根据数据分析结果，在宏观语言特征方面，郑译本相对于杨译本词汇更为丰富。在可读性检验方面，郑译本的语言更为复杂，阅读难度高于杨译本，适合初中 3 年级—高中 2 年级的读者群体阅读，杨译本是针对低年级学生优化的删减版本，适合初中 1—2 年级学生阅读，可读性计量分析对于出版社进行市场定位有一定的参考价值。此外，汉译本平均句长远远低于原著，而句子总数却远远超过原著，这说明汉译本普遍采用了分译的翻译方法。在主题性方面，汉译本的高频词呈现出高度一致性，均形成了特定的词汇聚类。然而，郑译本的词聚类联系更为紧密，对于原著主题信息和基调的还原更为密集和准确。在微观语言特征方面考察了连词和动词的十余种依存关系数据。连词计量分析的结果显示，汉译本的连词搭配主要表现为静态，而原著则更注重动态表达。此外，汉译本在连词的翻译上呈现显著的隐化现象，且存在着简化连接词搭配来降低阅读难度的趋势。在动词语言计量特征方面共 6 组数据。结果显示，汉译本更倾向于还原原著中隐化的名词主语成分，且动词拷贝结构的使用更频繁，同时相对较少使用动词—介词短语搭配，以增加信息表达的密度。此外，被动动词的使用比

例远低于原著，表明汉译本的被动态使用较少，多用主动态动词，符合汉语的表达习惯。

本研究也存在一些不足之处。首先，在计量特征的数量和维度方面还有待提高，未来的研究可以考虑增加计量特征的维度。其次，对于跨语种的计量特征检验方面仍有改进空间，未来的研究可以探讨跨语种的计量特征对比研究方法，以提供更加准确、全面和可靠的数据和研究经验，从数字人文的角度更好地研究文学原著和译本的语言特征。

附注

1 杨实译本由远方出版社于 1997 年出版，郑海凌译本由漓江出版社于 2006 年出版。

2 https://spacy.io

参考文献

[1]Baker Mona. Corpus Linguistics and Translation Studies, Implications and Applications[A]. Text and Technology: In Honor of John Sinclair[C]. Amsterdam: John Benjamins Publishing Company，1993.

[2]Baker Mona. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research [J]. Target, 1995(2).

[3]Baker Mona. Corpus-based Translation Studies: The Challenges that Lie Ahead [A] Terminology, LSP and Translation [C]. Amsterdam: John Benjamins Publishing Company, 1996.

[4]Gellerstam M. Translationese in Swedish Novels Translated from English [A]. Translation Studies in Scandinavia[C]. Lund: CWK Gleerup, 1986.

[5]Holmes J. S. The Name and Nature of Translation Studies[A]. Reader[C]. London: Routledge, 1972.

[6]傅琳凌，穆　雷. 语料库翻译学:在名与实之间[J]. 外语学刊，2020(1).

[7]何春艳，罗慧芳. 国内语料库翻译学研究动态的知识图谱分析(1993—2020)[J]. 中国科技翻译，2020(4).

[8]黄立波，朱志瑜. 语料库翻译学：研究对象与研究方法[J]. 中国外语，2012(6).

[9]刘国兵,常芳玲. 基于 CiteSpace 的国内语料库翻译学研究知识图谱分析[J]. 河南师范大学学报(自然科学版)，2018(6).

[10]王克非，黄立波. 语料库翻译学的几个术语[J]. 四川外语学院学报，2007(6).

[11]原　伟，刘海涛. 英俄语虚假新闻共性计量特征挖掘与跨语言聚类研究[J]. 数据分析与知识发现，2023(8).

[12]杨　子. 翻译构式观与语料库翻译学下的译者风格研究[J]. 上海翻译，2016(3).

Comparative Analysis of Quantitative Measurements on the Linguistic Features of the Chinese Translations of *Childhood*

Na Jun-yuan

(Heilongjiang University, Harbin 150080, China)

Abstract: Taking digital humanities as an entry point, this study adopts the corpus research paradigm to conduct a comparative study on the linguistic features of Chinese translations of *Childhood*, so as to provide a reference for the exploration of new paths for the study of Chinese translations of literary works.

A corpus is established by collecting the original text and the Chinese translation corpus, analyzing the macro and micro features of the Russian-Chinese corpus, and then exploring the similarities and differences of the linguistic features of the Chinese translation. The data results show that the Chinese translations present a high degree of similarity in terms of the measurement characteristics of the collocation paradigms of conjunctions and verbs, indicating that the gap in their linguistic dependency structure is relatively small; they present certain differences in terms of lexical richness, readability and thematicity tests, indicating that there is a large gap in linguistic complexity, readability and thematic expression between the Chinese and the Chinese translations. Compared with the source text, the Chinese translation shows a general tendency of implicit translation of conjunctions. In terms of the measurement characteristics of verbs, the verb-preposition phrase collocation of the Chinese translation shows a tendency of stealthy translation, and the number of verb copy structures is significantly larger than that of the source text, and the Chinese translation pays more attention to conveying the meaning of the verb itself, whereas the source text focuses on the collocation of the verb with other words to convey the meaning.

Keywords: *Childhood*; Chinese translation; linguistic features; quantitative measurements; comparative analysis

基金项目：本文系译国译民翻译研究院“新时代翻译与翻译行业研究课题“重点项目“数字人文视域下的视频游戏本地化翻译研究”（项目编号：Z202304）、黑龙江省高校智库开放课“东北抗战翻译资料整理与应用研究”（项目编号：ZKKF2022037）、黑龙江省教育科学规划重点课题“‘十四五’黑龙江省高校一流翻译学科建设探索与实践”（项目编号：GJB1423248）、黑龙江省学位与研究生教育教学改革研究项目“学科专业一体化高端翻译人才培养模式研究”（项目编号：JGXM__YJS__2022010）的阶段性成果。

作者简介：关秀娟（1975—），辽宁沈阳人，黑龙江大学俄语学院教授，研究方向：翻译学、俄罗斯汉学与中俄文明互鉴、俄汉对比。那峻源（1998—），黑龙江哈尔滨人，黑龙江大学俄语学院2021级硕士研究生，研究方向：翻译学。

收稿日期：2023-11-10 **[责任编辑：信　娜]**