

基于平行语料库的机器翻译中 12 类叠词研究

祁玉玲

(西安交通大学 外国语学院, 西安 710049)

摘要: 基于《傲慢与偏见》一对五平行语料库, 以百度在线翻译的《傲慢与偏见》为重点, 选取叠词作为研究对象, 进行百度在线翻译与人工译本在 12 类叠词翻译上异同的对比, 以及百度在线翻译存在的规律性特征的探寻。结果发现, 百度在线翻译的 12 类叠词与人工译本在 5 类叠词上尚存在显著性差异, 叠词的很多词类存在百度在线翻译中缺失的情况。虽与人工翻译存在一定的差距, 但研究也发现百度在线翻译的大部分叠词符合自然语言的分布规律。

关键词: 百度在线翻译; 人工译本; 叠词;

中图分类号: H315.9 **文献标识码:** A

A Parallel Corpus-based Study of 12 Types of Reduplicated Words in Machine Translation

QI Yuling

(Department of Foreign Studies, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Based upon the parallel corpus of five Chinese translations of *Pride and Prejudice*, especially the Baidu online translation as the study material and reduplicated words as the study focus, this study compared the human translation and machine translation and found out the regularities of Baidu online translation from the perspective of 12 types of reduplicated words. The results shows that the Baidu online translation varies significantly from human translation in 5 types of reduplicated words. Although the differences still exists, most of the reduplicated words translated by Baidu online follows the distribution of the natural language.

Key words: Baidu online translation; human translation; reduplicated words

1 引言

目前, 机器翻译已形成基于规则、基于统计以及基于实例的翻译方法, 虽目前很多人对机器翻译的结果尚抱有怀疑与不满的态度, 但仍不可否认机器翻译取得的进步和为翻译带来的便捷。机器翻译是建立在语言学、数学、信息学、计算机科学等学科基础上的边缘学科^[1]。从语言学的角度对机器翻译的研究, 不乏其例。蒋跃^[2]通过计量语言学方法, 对人工译本和机器译本的 13 个语言结构计量特征进行了考察, 从而对人工译本和在线机译译本的词法和句法层面的各种语言计量特征以及总体的翻译做出评比。祁玉玲等^[3]通过确立的 23 个语言计量特征建立起翻译质量评估模型, 对机器在线翻译进行了测评, 发现机器翻译在代词、副词、数词、量词、连词、助词和习语等词汇层面以及各类标点符号的比例等方面和人工译本差异较大。综上所述, 可发现从语言学角

度对机器翻译展开的研究, 可以对机器翻译的结果进行更为细致的评价, 从语言学的视角为提升机器翻译的质量提供一些建议。

本篇研究拟从叠词翻译的角度对百度在线翻译的具体结果进行探讨。叠词是汉语的一种特殊词汇现象, 使用非常普遍, 汉语的名词、数词、量词、形容词、副词、动词以及象声词都有重叠变化^[4]。这种叠字结构有二字格的、有三字格的, 也有四字格的。作为一个词, 重叠形态在意义上代表着一个特定的概念, 语音形式固定, 一般不能自由拆开^[5]。所以叠词的翻译可以很好的反映出翻译质量, 叠词使用数量以及种类的多寡, 也可作为衡量翻译质量的一个方面。

本研究以百度在线翻译为机器翻译的一个代表, 探讨机器翻译对文学作品的翻译情况, 选择的切入点就是不同类型的叠词的翻译情况, 重点通过对比百度在线翻译与人工译者具体在叠词翻译方面的异同, 探讨如何更好的改进以百度翻译为代表的机器翻

译, 使之更好的帮助人工翻译, 更好的实现机助人译或者人机互助模式。

1. 百度在线翻译和人工译本在翻译的叠词方面是否存在差别? 如果答案是肯定的, 具体存在何种类型的差别, 显著与否?
2. 百度在线翻译的叠词自身是否存在某些规律性的特征?
3. 百度翻译和其他的机器翻译有多大的差别?

2 研究方法

2.1 研究语料

本研究基于自建的英汉双语一对五平行语料库, 即一个英语原文《傲慢与偏见》对应 5 个汉语译文, 汉语译文来自四个人工译者和一个百度在线翻译。其中, 英语原文在 12 万字左右, 而每个汉语译文在 20 万字左右。为了后续的研究考虑, 已经对语料进行过人工清洗、标注和对齐等操作。

2.2 研究设计

在以往研究的基础上^[6], 我们对叠词进行了归类, 分成 12 类。而若要在软件中, 实现对叠词的批量提取, 则需要借助正则表达式, 在借鉴秦洪武等^[7]有关正则表达式的基础上, 我们对每一类叠词的正则表达式进行了编写, 在 Antconc 中每一类叠词的正则表达式, 如下表所示:

表 1 Antconc 中 12 类叠词的正则表达式

叠词类型	正则表达式
AA	$\backslash\mathbf{b}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{2}\backslash\mathbf{b}$
ABB	$\backslash\mathbf{b}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{3}\backslash\mathbf{b}$
AAB	$\backslash\mathbf{b}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{2}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{b}$
ABA	$\backslash\mathbf{b}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{2}\backslash\mathbf{b}$
AABB	$\backslash\mathbf{b}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{2}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{3}/\backslash\mathbf{b}$
BCAA	$\backslash\mathbf{b}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{4}/\backslash\mathbf{b}$
ABAC	$\backslash\mathbf{b}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{2}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{b}$
AABC	$\backslash\mathbf{b}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{2}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{b}$

ABAB	$\backslash\mathbf{b}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{2}/\backslash\mathbf{b}$ 或 $\backslash\mathbf{b}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{2}\backslash\mathbf{3}/\backslash\mathbf{b}$
ABCA	$\backslash\mathbf{b}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{2}/[\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}]+\backslash\mathbf{b}$
BAAC	$\backslash\mathbf{b}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{3}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{b}$
BACA	$\backslash\mathbf{b}([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])([\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}])\backslash\mathbf{3}/[\backslash\mathbf{x}00-\backslash\mathbf{x}\mathbf{f}\mathbf{f}]+\backslash\mathbf{b}$

借助上表中所列的正则表达式, 在 Antconc 中对百度在线译本和人工译本中出现的每一类叠词进行提取。

3 结果与讨论

3.1 叠词的形符和类符数

提取过程中发现, 12 类叠词在五个译本中只出现了 11 类, ABAB 类叠词并未检索到结果, 笔者猜测可能是由于分词软件在自动分词时, 将 ABAB 类叠词自动分成 AB 和 AB 导致的结果, 所以此类叠词暂不在我们研究之列。为了很好地对百度在线翻译和人工翻译进行对比, 提取之后的形符和类符, 进行了卡方检验, 看二者是否存在显著性差异。百度译本和四个人工译本中 11 类叠词的类符和形符如表 2 和表 3 所示。

表 2 11 类叠词的类符对比

叠词类型	百度	人工	人/机
AA	63	127.5	0.000 ***
ABB	5	33.75	0.000 ***
AAB	7	11.5	0.295
ABA	6	8.75	0.474
AABB	8	64.25	0.000 ***
BCAA	8	16.25	0.094
ABAC	16	76.75	0.000 ***
AABC	9	34.5	0.000 ***
ABCA	1	6.5	0.045 *
BAAC	0	0.75	0.386
BACA	2	8.75	0.040 *
类符总数	117	389.25	0.000***

注: 人工指的是四个人工译本的平均值

表 3 11 类叠词的形符对比

叠词类型	百度	人工	人/机
------	----	----	-----

AA	922	1471.25	0.000 ***
ABB	6	61.25	0.000***
AAB	16	18.5	0.67
ABA	44	50.5	0.504
AABB	13	131.25	0.000 ***
BCAA	9	37.5	0.000***
ABAC	22	157.5	0.000 ***
AABC	11	68.75	0.000 ***
ABCA	1	9	0.011 *
BAAC	0	1.25	0.264
BACA	2	14	0.003 **
形符总数	1037	2020.75	0.000 ***

注：人工指的是四个人工译本的平均值

总体来看，百度在线译本在 11 类叠词的类符和形符上都较低，说明百度在线翻译在叠词的丰富度和叠词的使用频次上还存在较大差别。百度和人工译本在 5 类叠词的形符和类符上都存在显著性差异，分别是 AA 式叠词、ABB 式叠词、AABB 式叠词、ABAC 式叠词和 AABC 式叠词。

表 4 百度在线翻译的 11 类叠词的词类分析

叠词分类	百度在线译本	人工译本	缺少的词类	未出现在百度中的叠词举例
AA	副词，动词，名词，状态词，量词，形容词，叹词	副词，动词，名词，状态词，量词，形容词，代词，语气词，区别词，叹词，数词，拟声词	代词，语气词，区别词，数词，拟声词	哈哈，偏偏，频频，漫漫等
ABB	状态词，数词，副词，名词	状态词，数词，副词，名词，语气词	语气词	兴冲冲，乐滋滋，冷冰冰，醉醺醺，阴沉沉等
AAB	状态词，名词	状态词，动词，名词，形容词	动词，形容词	散散步，点点头，飘飘然，鼓鼓的，献献丑等
ABA	副词，动词	副词，动词，名词	名词	时不时，越来越，动不动等
AABB	状态词，动词	状态词，副词，动词，形容词，名词，量词，区别词，拟声词，时间词，方位词	副词，形容词，名词，量词，区别词，拟声词，时间词，方位词	原原本本，口口声声，疯疯癫癫，絮絮叨叨等

由于叠词包含了译者的主观感情色彩，所以人工译者在翻译时会结合语境与人物的情感，将一些简单的形容词，用叠词来加以描述，增强语言的表达效果与感染力。而百度在线翻译，对于一些简单的形容词会更倾向于找一些相对应的中文意思进行匹配，这可能与百度在线翻译所接触的学习语料有关，建议可在后续的机器学习中，加入一些叠词的英汉短语对或句子对，可为机器翻译在进行翻译活动时提供更多的参考和选择。另外，我们也欣喜的发现，AAB、ABA 和 BAAC 式三类叠词，无论在形符和类符上，百度在线翻译和人工译本都不存在差异，在某些叠词的类符上，两类翻译差异也不明显，所以百度在线翻译也是有可圈可点之处的。

3.2 词类分析

11 类叠词中，每一类叠词都有不同的词类，对比每一类叠词的词类构成，可发现百度在线翻译在翻译成叠词的过程中，会倾向于翻译成哪些词类的叠词，而哪些词类还存在欠缺。结果如表 4 所示，

BCAA	动词, 形容词, 副词	动词, 形容词, 副词, 名词, 状态词	名词, 状态词	忧心忡忡, 气势汹汹, 风度翩翩, 可怜巴巴等
ABAC	动词, 形容词, 副词, 名词	动词, 形容词, 副词, 名词, 区别词, 状态词, 量词	区别词, 状态词, 量词	合情合理, 不知不觉, 自由自在, 非亲非故等
AABC	动词, 形容词, 名词	动词, 形容词, 名词, 副词, 连词, 状态词	副词, 连词, 状态词	咄咄逼人, 姗姗来迟, 落落大方, 念念不忘等
ABCA	动词, 量词	动词, 形容词, 区别词, 量词, 名词	形容词, 区别词, 名词	忍无可忍, 闻所未闻, 日复一日, 数不胜数等
BAAC	无	动词	动词	自欺欺人
BACA	形容词, 副词	动词, 形容词, 副词, 名词, 数词	动词, 名词, 数词	以讹传讹, 听之任之, 得过且过, 自然而然等

通过观察上表可发现, 百度在线译本中叠词的词类相较于人工译本的词类, 要少一些, 这是由于人工译者在翻译的时候, 会根据主观理解, 改变英语原文中词的词性, 使调整后的词性和词义更适于表达原文的意思。由于百度在线翻译是基于深度学习, 通过学习大量双语互译对照的句子, 逐渐提升翻译水平, 那是否一方面是由于此类文学方面学习资料的欠缺, 另一方面是由于文学作品的翻译, 人工译者在自身理解的基础上, 加入了情感, 而百度在线翻译, 目前, 只能做到“达意”, “表情”还是一个不小的挑战。

3.3 百度在线翻译的叠词的规律性特征

表 5 百度在线译本中 AA 类叠词的使用情况

序列	频率	词语	序列	频率	词语	序列	频率	词语
1	306	太太	22	5	听听	43	1	偏偏
2	135	妹妹	23	4	啊啊	44	1	冷冷
3	102	姐姐	24	4	微微	45	1	婆婆
4	69	叔叔	25	4	舅舅	46	1	学学
5	40	哥哥	26	3	一一	47	1	小小
6	30	妈妈	27	3	万万	48	1	开开
7	22	常常	28	3	悄悄	49	1	摇摇
8	18	看看	29	3	慢慢	50	1	整整
9	17	姑姑	30	2	乖乖	51	1	最最

为了更好的对百度的翻译语言进行研究, 我们将采用计量语言学的一些方法, 对百度翻译语言加以分析。关于百度翻译中的每一类叠词的使用, 是否符合一定的自然语言规律, 将用 Zipf 定律加以验证。Zipf 定律是研究频率词典中单词的序号与频度的分布规律^[8]。作为计量语言学的第一定律, Zipf 定律是一个在计量语言学中具有核心地位、与诸多语言性质和语言过程密切相关的语言规律^[9], Zipf 定律指出, 每个从自然文本得到的这种词语序列中的词频与他们所对应的序列之间呈现反比例关系^[10]。百度翻译语言是否符合自然语言的分布规律呢, 我们将用提取的数据进行进一步的探讨。

10	14	往往	31	2	匆匆	52	1	望望
11	13	弟弟	32	2	好好	53	1	样样
12	12	刚刚	33	2	想想	54	1	步步
13	11	谢谢	34	2	明明	55	1	点点
14	10	仅仅	35	2	深深	56	1	狠狠
15	8	大大	36	2	渐渐	57	1	短短
16	8	婶婶	37	2	试试	58	1	种种
17	8	爸爸	38	2	轻轻	59	1	胖胖
18	8	等等	39	2	静静	60	1	见见
19	6	某某	40	2	默默	61	1	谈谈
20	6	远远	41	1	个个	62	1	走走
21	5	匆匆	42	1	人人	63	1	问问

如上表所示，频率最高的是“太太”一词，频率达到了 306 次，剩余的 62 个词语频次依次递减，频率高于 100 的仅有 3 个词，而频率低于 5 的有 41 个词语，仅出现 1 次的就有 23 个，这种词语的分布现象，即频率高的词语，数量少，频率低的词语，却数量较多，有多大程度上符合语言的分布规律呢。我们将用 Zipf 定律来验证，定律的公式如下：

$$y=ax^{-b}$$

变量 y 表示词语的频率，变量 x 表示词语的序列号， a 、 b 是两个参数。拟合情况如下：

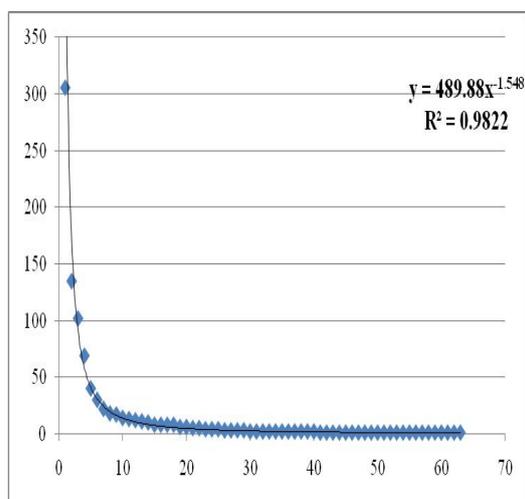


图 1 AA 式叠词频率分布曲线及 Zipf 拟合图

用公式拟合 AA 式叠词词频序列中的“序列—频率”两列数据得到的 $a=489.88$ ， $b=1.548$ ， R^2 值为 0.9822，大于 0.75，说明百度在线翻译译本中 AA 式叠词的词频分布是符合 Zipf 定律的。图中一个个散点对应词频序列中的一个词，即有 63 个点，而实线则是拟合曲线，表示理论上的数值。剩下的 10 类叠词的数据和拟合曲线，由于篇幅问题，就不一一列出，但将对 R^2 值有一个整体的呈现。

表 6 各类叠词与 Zipf 定律的拟合情况

叠词	ABB	AAB	ABA	AAB B	BCA A
R^2	0.709	0.916	0.933	0.889	0.579
	4	6	3	6	8
叠词	ABA C	AAB C	BAA C	ABC A	BAC A
R^2	0.800	0.719	无	0.776	无
	5	1		3	

可以发现，AAB 式、ABA 式、AABB 式、ABAC 式和 ABCA 式 5 类叠词的拟合优度 R^2 值大于 0.75，而 ABB 式、BCAA 式和 AABC 式没能很好的和 Zipf 曲线拟合。为此，我们探究了人工译本对此三类叠词的拟合情况。结果发现，人工译本的 ABB 式叠词、BCAA 式叠词和 AABC 式叠词 R^2 值都在 0.75 以上，能很好的和 Zipf 曲线拟合。也许，百度在线翻译应在进行语料学习的时

候,适当的增加此类叠词的学习,无论是英译汉的语料亦或是汉译英的语料。BAAC 式叠词,由于百度在线翻译中未检索到此类叠词, BACA 式叠词只有 2 个,所以,无法进行拟合。

3.4 百度在线翻译中词类的分布情况

前文提到,百度在线翻译不仅在各类叠词的类符和形符上远远低于人工译本,而且词类也有所缺乏。而百度在线翻译的词类分布现象是否是符合规律的呢,将用 Zipf-Alekseev 定律进行进一步的探究。此定律是 Zipf 定律的一种变体,同时也是计量语言学中经常使用的一种频率分布模型,这个模型遵循一个原则:语言单位或语言结构是多样化的(diversification),所以这个模型能适用于多种语言层面。例如词频分布,句法关系分布,话语关系分布等^[11]。该定律的公式如下^[12]:

$$P_x = \begin{cases} \alpha, & x=1 \\ \frac{(1-\alpha)x^{(a+b\ln x)}}{T}, & x=2,3,\dots,(n) \end{cases}$$

$$T = \sum_{j=2}^n j^{-(a+b\ln j)}, 0 < \alpha < 1, a, b \in R$$

拟合后的结果如下图所示:

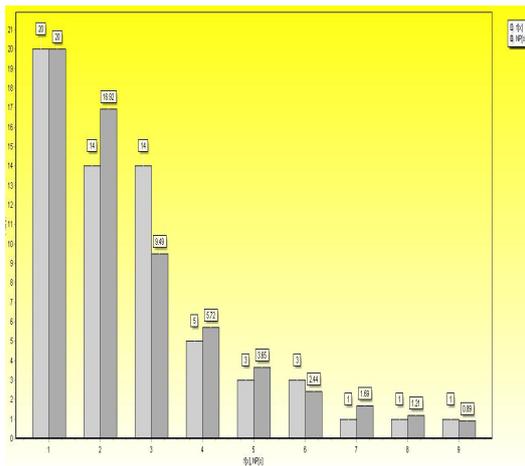


图 2 AA 式叠词的词类分布曲线及 Zipf-Alekseev 定律拟合图

图中 $f[x]$ 表示百度在线翻译中的数值, $NP[x]$ 表示理论上的值, 其中 $a=0.5514$, $b=0.4878$, $n=9$, $\alpha =0.3226$, $DF=3$, $R^2 =0.9267$ 。 R^2 的值由于大于 0.75, 说明很好

的和 Zipf-Alekseev 定律实现了拟合。所以结果表明,百度在线翻译中 AA 式叠词的翻译有一定的分布规律,此规律是和自然语言的分布规律相一致。

剩余的 10 个叠词词类,由于词类较少,不能进行拟合实验,所以将不再进行讨论。

就叠词来说,百度翻译确与人工翻译之间存在明显的差异,同时,自身也存在一定的规律性特征,而百度翻译是否就代表了机器翻译呢?还需要进一步的验证。

3.5 百度翻译和其他的机器翻译的差别

为了验证百度翻译和其他的机器翻译之间并无明显差异,我们选取有道和谷歌翻译同百度翻译进行比较。依然是以《傲慢与偏见》为例,选取了其中的 6 个章节,分别是 1、11、21、31、41、51 和 61,然后对比三大翻译系统的结果,看彼此之间是否存在显著性差异。

表 7 三大翻译系统结果对比

	百度		有道		谷歌	
	类	形	类	形	类	形
AA	48	158	24	94	50	182
AAB	1	1	\	\	\	\
ABA	2	3	\	\	3	5
AABB	1	1	1	2	3	3
ABAC	1	1	1	1	4	4
BCAA	2	2	\	\	\	\
ABCA	\	\	\	\	1	1

通过对翻译结果检索发现,翻译结果中只有 7 类叠词,其他的 5 类叠词并未出现。而在这 7 类叠词中,AA 类叠词数量最多,通过对 AA 类叠词进行卡方检验,发现百度和谷歌之间并没有显著性差异,而有道和这两者在 AA 类的叠词上是存在显著性差异的。除 AA 类叠词外,其他类的叠词数量普遍较少,并无显著性差别。以此得出结论,虽在一些细节上存在差别,但百度翻译基本上可以代表机器翻译。

4 结语

1. 百度在线翻译在英译汉的文学作品翻译中,虽在某些类叠词的使用频次及叠词的丰富度上还与人工译本存在差别,但某

些叠词的翻译也可与人工相比拟。在今后，可以在对机器翻译的训练中，增加一些叠词的短语对和句子对的学习，相信机器翻译可以逐步缩小与人工翻译之间的差距，更好的为人工翻译服务。

2. 百度在线翻译在某些叠词的词类上与人工翻译存在一定的差别，存在有些词类缺失的情况。这与结论 1 的建议一样，增加有关叠词方面的输入。

3. 百度在线翻译中的词频分布与词类分布都符合自然语言的分布规律，这足以说明百度在线翻译的翻译语言已在很大程度上接近自然语言。

4. 百度翻译的结果和谷歌等其他在线翻译软件之间，在叠词的翻译上并无显著性的差异。

参考文献

- [1] 戴新宇, 尹存燕, 陈家骏, 等. 机器翻译研究现状与展望[J]. 计算机科学, 2004, 31 (11): 176-179.
- [2] 蒋跃. 人工译本与机器在线译本的语言计量特征对比—以 5 届韩素音翻译竞赛英译汉人工译本和在线译本为例[J]. 外语教学, 2014, 35 (5): 98-102.
- [3] 祁玉玲, 蒋跃. 基于语言计量特征的文学翻译质量评估模型的构建[J]. 西安电子科技大学学报(社会科学版), 2016, 26 (1): 84-92.
- [4] 周笃宝. 汉语叠词与翻译[J]. 中国翻译, 1999, 03: 27-29.
- [5] 汪国萍. 汉语叠词和英语词汇重叠现象对比分析[J]. 浙江师范大学学报, 2003, 28(4): 31-34.
- [6] 郭志良. 有关“AABB”重叠式的几个问题[J]. 语言教学与研究, 1987, 02: 60-70.
- [7] 秦洪武, 王克非. 正则表达式在汉语语料检索中的应用[J]. 外国语文, 2013, 29 (6): 74-79.
- [8] 冯志伟. 用计量方法研究语言[J]. 外语教学与研究, 2012, 44 (2): 256-269.
- [9] 刘海涛, 黄伟. 计量语言学的现状、理论与方法[J]. 计算机工程与应用, 2012, 42 (2): 178-192.
- [10] ZIPF, G. The Psycho-Biology of Language: An Introduction to Dynamic Philology[M] London: George Routledge & Sons Ltd. ,

1936:152-224.

- [11] 刘海涛, 潘夏星. 汉语新诗的计量特征[J]. 山西大学学报(哲学社会科学版), 2015, 38 (2): 40-47.
- [12] Strauss, U. & Altmann, G. Diversification Laws in Quantitative Linguistics. [EB]. [2006].<http://lql.uni-trier.de/index.php/Diversification>.



祁玉玲 (1991-), 硕士, 主要研究领域为语料库翻译学。
Email: 15229212810@163.com