目录

第一章　绪　论

1.1　研究背景

国内外英语水平考试层出不穷，已经形成了不同体系。国内有CET、PETS、TEM、研究生入学考试等各级各类英语水平考试，国外有TOEFL、GRE、IELTS等语言水平测试。这些考试无一例外都包含主观题，例如，在托福IBT测试中，听说读写四个部分基本上以主观题为主要测试形式，主观题测试已经成为衡量学习者外语水平的重要标尺。作为五种英语技能之一的翻译，在国内的上述主观题测试中出现频率也颇高，例如，在考生人数众多的四、六级考试中，改革前是单句汉译英，改革后调整为段落汉译英，翻译内容涉及中国的历史、文化、经济、社会发展等各方面。四级汉译英段落包含140—160字，六级段落长度为180—200字（参见四、六级考试委员会网站http://www.cet.edu.cn/slj.htm）。此外，我国还有针对翻译技能的专门测试，如全国翻译专业资格考试（CATTI）、全国外语翻译证书考试（NAETI）、上海外语口译证书考试等。

全国翻译专业资格考试（CATTI）由我国外文局翻译专业资格考试中心举办，首次考试时间为2003年，考生人数仅为1682人；2021年报考量突破35万人；2022年仅上半年，考生报名人数近18万，呈现几何级数增长（详见http://hebei.ifeng.com/c/8RmzzYuUpTu）。每年参加英语专业八级考试的人数也在万人以上，而大学英语四、六级考试的参加人数更为惊人，2023年四、六级考试全国报名人数为2035万人（详见https://m.sxks114.com/cn/h-nd-12890.html）。如此庞大的考生队伍必然给阅卷工作带来繁重的负担，为了减轻阅卷人员的工作量，提高阅卷效率，翻译自动评分系统的研制势在必行。

其次，翻译测试评分是主观性比较强的项目，对评分员要求比较高。由于考生众多，必然需要邀请数量可观的评分员。首先，找到如此众多合格评分员实属不易。其次，在高强度的评分过程中要求如此众多的评分员一以贯之地把握评分标准更是难上加难。如果评分质量受到限制，势必影响评分的信度和效度。此外，低质量的评分也会给考生带来不良影响，甚至有可能影响他们的人生轨迹。Zhang（2013: 2）在 ETS 的报告中曾指出人工评分的三个不足："首先，必须要聘请到合格的评分员；其次，必须要培训评分员如何把握评分标准，并在正式评分前保证他们具有合格的评分能力；最后，必须要在评分过程中密切监控评分员，以保证评分质量和一致性。"总之，在大规模考试中保证人工评分的信度和效度是一项巨大的工程，其结果往往不太尽如人意。

再者，计算机技术的飞速发展为人工评分转为机器自动评分带来了十分难得的机遇。国内外相继问世的计算机自动评分系统尤其是英语作文评分系统，对汉译英自动评分系统的研究与构建具有较强的借鉴意义。

计算机自动评分系统的应用不仅能够节约成本（Bereiter 2003; Chung & O'Neil 1997; Page 2003），还可以有效地提高评阅的一致性和效率。Mason（2002）指出，英国教师 30% 的时间都花在了评分上，而这 30% 时间的价值是 30 亿英镑。通过计算机评分可以减轻教师工作量，提高工作效率，减少不必要的经费开支。David et al.（2010: 1）指出，"计算机评分系统与人工评分相比，速度更快，效率更高。此外，自动评分系统一致性更高，更为公平，也便于对考生成绩做长期跟踪分析，有些评分系统还能够对考生在考试中的表现提供细致的反馈"。Page（2003: 46）在"Project Essay Grade: PEG"一文中指出，"自动评分的准确性一般要高于两个评分员之间的评分信度（准确性指与评分员均分的一致程度）"。国内外作文自动评分系统的研究都表明机器评分与人工评分之间的一致性比较高（Attali 2004; Burstein & Chodorow 1999; Elliot 2000, 2002, 2003; Landauer et al. 2003; Nichols 2004; 梁茂成 2005; Wang & Stallone 2008; Shermis 2014）。

国内外学界在作文自动评分系统方面研究较为深入，而在翻译自动评价方面相对不足。作文自动评价和翻译自动评价虽有区别，但有很多

共性，作文自动评分系统在架构、模块、变量、研制方法等方面可以为翻译自动评分系统提供有益的思考和借鉴。

本研究是王金铨（2008）的后续研究，基于对英语专业高年级学生300篇叙事文汉译英文本的分析，利用多种语料库检索技术、信息挖掘技术、自然语言处理技术，从训练集译文中提取多个能反映学生译文质量的文本特征，通过回归分析构建了叙事文汉译英篇章、单句的诊断性评分模型，和适用于大规模测试的选拔性评分模型，并对同题汉译英验证集文本进行自动评分。研究结果表明，初步评分模型表现良好，能够比较准确地预测中国二语学习者的汉译英成绩。但是，王金铨（2008）的研究只是一个开端，以一篇叙事文为语料进行了理论探索和实践创新，初步构建了汉译英评分系统模型。不过，对于汉译英评分系统而言，基于一篇汉译英文本构建的评分模型还有待进一步验证，在文章体裁、文本数量等方面进一步扩充，最终形成预测力强、运行稳定、适应多种体裁的自动评分系统。

1.2 研究目的及研究问题

本研究将综合英语写作自动评分和机器翻译自动评价两个领域的知识，利用语料库语言学、信息检索、统计学以及自然语言处理领域中的相关知识，通过提取学生译作中与译文质量相关的多种文本特征，进行多元回归分析，构建有效适应不同文体的汉译英自动评分模型，既可以用于日常汉译英训练，也可以应用于大规模翻译测试评分。

本书的研究目的有三方面：

（1）探索适合中国二语学习者的汉译英机助评分系统。虽然国内外作文自动评分系统的研究较多，但翻译自动评分和作文自动评分区别比较大。首先，翻译有原文限制，在内容上不会像作文那样享有较大的自由度，字数也应在一定范围之内；其次，翻译更注重对意义的传达。奈达（Nida & Taber 1969: 12）将翻译定义为“在译语中用最切近而又自然的对等语再现原语信息，首先是意义，其次是文体”。意义的重要性超过

形式，甚至在某些情况下，为忠实传达意义而牺牲形式。翻译与写作的不同特点决定了翻译自动评分系统一定会具有不同于作文评分系统的特点和元素，这也是本研究的重点和翻译自动评分系统成败的关键。

（2）挖掘能够预测中国二语学习者汉译英不同文体译文的有效变量。作文自动评分系统能够为翻译自动评分系统提供比较好的借鉴作用，但是作文系统中现有变量对不同文体汉译英文本的预测能力尚未证实；其次，鉴于翻译自身的特点，需要挖掘更多符合汉译英特点的有效预测变量。

（3）在更大文本范围和更多文体中验证王金铨（2008）所构建的评分模型。王金铨（2008）的研究只限于一种文体（叙事文），且只一次性做了一种比例（150 训练集：150 验证集）的诊断性测试评分模型和四种比例（30：270；50：250；100：200；150：150）的选拔性测试评分模型的构建和验证工作，诊断性测试评分模型包括形式评分模型和语义评分模型，选拔性测试评分模型只包含语义模型。研究结果表明由 100 篇译文构成的训练集在评分信度和效度上能够满足译文自动评分系统的需要。本研究将在三种文体、多种比例训练集（100 篇及以上训练集）、多次随机的基础上对王金铨（2008）的研究做进一步验证和拓展，反复尝试，力争挖掘更多、更有预测力的文本变量，不断优化汉译英测试自动评分模型。

本书试图解决的研究问题如下：

问题一：汉译英自动评分系统中有效预测变量有哪些？是否对不同文体译文都有效？

问题二：译文质量预测因子构建的模型在不同文体中的预测能力如何？汉译英自动评分系统的评分信度能否达到语言测试的要求？

问题三：汉译英自动评分系统对不同文体译文进行评分时是否具有同等效果？训练集译文的最低样本量至少应该达到多少？

1.3 研究概述

王金铨（2008）的研究以一篇叙事文汉译英文本为语料创建了中国学生汉译英自动评分模型。本书在此基础上进一步拓展，涵盖了叙事文、说明文和议论文三种文体，包括近千篇文本，在多种文体、更大语料规模和更多训练集比例的基础上验证初始评分模型的预测力和稳定性。由于单句翻译缺乏上下文，不利于整体把握和理解，且翻译测试中的单句翻译近乎绝迹，原本包含单句翻译的四级考试也进行了改革，全国大学英语四、六级考委会已于 2013 年 8 月将单句汉译英调整为段落汉译英测试。另外，从翻译实践来看，有上下文的篇章翻译更符合阅读习惯，也更有利于文本理解和翻译。王金铨（2008）的研究表明篇章译文评分模型中所使用的预测变量完全适用于单句译文自动评分，本研究将以汉译英篇章为基础进行多种文体评分模型构建。

本研究历经了语料收集、语料转写、人工评分、变量挖掘、初始模型创建、初始模型验证、更大规模模型创建和验证等几个重要阶段。

在语料收集阶段，本研究收集了来自国内 11 个省市自治区、18 所不同水平层次大学英语专业三、四年级的汉译英语料，涵盖了 985、211、地方综合性大学、外语院校、理工院校、师范院校等多层次、多类型的高校，尽可能使语料更具代表性和广泛性。

在语料转写阶段，所有纸质文本形式的语料都被逐篇录入电脑，录入过程中最大限度地保留了学生译文的原貌，包括各种错误。

在人工评分阶段，由于本研究首先创建的是叙事文汉译英评分模型，研究者制定了详细的形式和语义评分细则，组织了多名有丰富英汉翻译阅卷经验的评分员对叙事文语料进行逐句评分，包括语义评分和形式评分。后续说明文和议论文人工评分将继续沿用叙事文评分标准，以保证人工评分的一致性。

在变量挖掘阶段，研究团队利用语料库分析技术、信息检索技术和自然语言处理等技术从译文文本中提取了大量能够预测译文质量的文本变量，包括形式变量和语义变量，分别作为三种文体形式和语义评分模型的自变量。

各文体语料将进行多次随机分组，形成不同比例的训练集和验证集，训练集用来构建各文体自动评分模型，验证集用来验证所构建自动评分模型的性能和预测力。

在模型构建阶段，本研究利用自然语言处理技术、信息检索技术和语料库分析技术，从三种文体训练集汉译英文本中提取多个能反映学生译作质量的文本特征项作为自变量，然后通过统计分析，计算自变量与因变量（人工评分）之间的相关系数，确定进入评分模型的预测因子，以相应形式或语义成绩作为因变量进行多元回归分析，反复尝试，不断优化，得到多元回归方程，该方程即可用来为同题译文进行自动评分。

在模型验证阶段，利用模型构建阶段所得到的统计模型为验证集译文自动评分，计算机器评分与人工评分之间的相关系数以确定初始模型的信度和效度。

为检验评分系统的稳定性和预测力，在模型构建阶段，本研究以100篇译文训练集为起点，以二分之一训练集为终点，通过五次随机的方式产生不同比例的训练集和验证集，构建自动评分模型，验证预测变量和回归模型的性能和稳定性，三种文体不同比例评分模型创建情况如下：

表 1.1　三种文体自动评分模型创建说明

文体	篇数	训练集验证集比例	随机分组次数	构建模型数量
叙事文	300	100 ： 200 110 ： 190 120 ： 180 130 ： 170 140 ： 160 150 ： 150	每个比例形式、语义评分模型随机分组 5 次	60
说明文	336	100 ： 236 110 ： 226 120 ： 216 130 ： 206 140 ： 196 150 ： 186 160 ： 176 170 ： 166	每个比例形式、语义评分模型随机分组 5 次	80

（待续）

（续表）

文体	篇数	训练集验证集比例	随机分组次数	构建模型数量
议论文	257	100 ： 157 110 ： 147 120 ： 137 130 ： 127	每个比例形式、语义评分模型随机分组 5 次	40

表 1.1 显示本研究将创建三种文体形式、语义自动评分模型共计 18 个比例，每个比例训练集、验证集随机分组 5 次，共创建 180 个自动评分模型。王金铨（2008）的研究只创建了叙事文一种比例（150 篇训练集：150 篇验证集）的诊断性测试评分模型和四种比例（30：270；50：250；100：200；150：150）的选拔性测试评分模型，且只进行了一次随机分组。研究结果显示，100 篇训练集和 150 篇训练集表现都很好，增加训练集后，模型性能提高有限，从提高人工评分效率出发，以 100 篇训练集构建的评分模型较为符合评分模型的需要。本研究在此结论基础上，以 100 篇训练集为起点，增加不同文体、训练集比例和随机分组次数，构建三种文体的汉译英自动评分模型，验证并拓展王金铨（2008）的研究结果，进一步完善中国学生汉译英评分系统的构建工作。

1.4 研究结构

本书共包含九章。

第一章介绍本研究的背景、研究目的和问题，以及对本研究进行概述。

第二章回顾分析了计算机辅助语言测试的理论基础和计算机技术在语言测试中的应用，并在此基础上指出计算机辅助语言测试所面临的挑战和发展方向。

第三章回顾近五十年来计算机评分系统的发展历程和主要特点，探讨现有自动评分系统的优缺点，对本研究中汉译英机器评分系统的框架、模块和预测变量提出总体设计和建设构想。

第四章介绍现有翻译质量评估方法及其在计算机自动评分系统中的应用，综合翻译自动评价系统和作文自动评分系统中对语言质量进行评

价的手段和方法，并提出适合中国学生汉译英自动评分系统的量化评测方法。

第五章介绍本研究中所使用的语料、汉译英评分标准的制定、评分过程和评分信度。

第六章介绍本研究中使用的自然语言处理工具，包括文本预处理工具、文本分析工具和数据分析工具，描述评分模型中形式、语义文本特征及其提取方法和过程，并对数据分析流程进行了概述。

第七章考察所提取的文本特征与三种文体译文质量之间的关系，遴选进入模型构建过程的译文质量预测因子。

第八章构建三种文体汉译英测试自动评分模型，考察评分模型的预测能力和评分信度，比较三种文体自动评分系统，考察预测变量的稳定性和预测力。

第九章为本研究的结论部分，总结本研究的主要发现，指出模型应用中的注意事项、研究价值、研究不足，并对将来的研究工作做一个系统回顾和展望。