分类号：_____ 单位代码：10033

密　级：_____ 学　　号：63110050102008

中国传媒大学

博士学位论文

汉维平行语料库构建研究

Study on the Construction of Chinese-Uyghur Parallel Corpus

申请人姓名：塞麦提•麦麦提敏

指导教师：侯敏　教授

专业名称：语言学及应用语言学

研究方向：计算语言学

所在学院：文学院

论文提交日期 2009 年 4 月

中国传媒大学研究生学位论文独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 **中国传媒大学** 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：　　签字日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解 **中国传媒大学** 有权保留并向国家有关部门或机构送交本论文的复印件和磁盘，允许论文被查阅和借阅。本人授权**中国传媒大学**可以将学位论文的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

致谢

在中国传媒大学的三年学习对我来说就像三天之内发生的事情。似乎记得大前天还在面试，在导师和其他老师的面前紧张得一塌糊涂，记得前天才到播音主持艺术学院报到，记得昨天还在侯老师家里谈论文应如何着手。怎么这么快就到今天了呢？！然而，这三天却是我人生中最重要的日子。在这些日子里有许多重要的事，许多重要的人将铭刻在我的记忆里永不会抹去。

三年难忘的博士生活即将结束，在此论文完成之际，我首先要衷心感谢我的导师侯敏教授三年来的严格要求、悉心指导和无私帮助。她对我博士论文的选题、研究、写作的指导注入了大量的精力。她严谨求实的治学态度、渊博的专业知识和创新开拓的精神以及对事业的执著、热情将永远是我学习的榜样。她的关怀我将永生难忘。

我要深深感谢冯志伟教授的关怀和指导。他高瞻远瞩的学术思想、平易近人的作风令人敬佩。他渊博的知识、宽广的胸怀、无私奉献的精神给了我深刻的影响。在攻读博士学位期间，我一直得到的他亲切关怀、鼓励和耐心教诲，在此，向他表示深深的敬意。

我要特别感谢力提甫教授、吐尔根教授、傅爱平研究员。他们评阅我的博士论文并提出了非常宝贵的建议和意见。

我还感谢中国传媒大学的刘海涛教授、邹煜博士、胡凤国博士、何伟博士，是他们在专业和生活上，毫无保留地给了我许多的帮助，他们的指导，让我少走了许多弯路。

我还要感谢中国传媒大学有声媒体语言监测和研究分中心的所有老师和同学，平时跟他们相互交流学术观点，让我受益非浅。感谢他们的直接或间接帮助。

我要感谢所有给予过我知识、关怀过我的老师们。至今我还清晰地记着，木哈拜提教授把我引入语言学的殿堂，之后又指导我完成硕士学位论文。

新疆大学人文学院是我工作的单位，也是我坚强的后盾，这里的领导、同事以及朋友都给了我很大的支持。在此向他们表示衷心的感谢。

最后我还要衷心感谢我的父母和亲戚，没有他们在精神上的鼓励和生活上的支持，我不可能顺利地完成学业。

感谢我爱人，她不仅能充分地理解和体谅我学习的辛苦，而且在生活上承担大部分的家务，鼓励我安心学习，她的支持和帮助使我顺利完成论文。

我是一个非常幸福、非常幸运的人。在我周围，还有很多的领导、老师、朋友、熟人、亲戚，他们都为我的学习提供了很大的帮助。最后，我要深深感谢所有那些给我关心和支持的所有人。

摘　要

平行语料库（Parallel corpora）是由原语文本及其平行对应的译语文本构成的语料库。创建平行语料库的主要目标是以计算机储存大量的两种语言的对应语料，对语料做各种标注，利用研制的检索工具进行快捷的检索和分析，以发现各种语言现象和双语之间的对译信息，并为语言研究和其它领域研究提供提取这些信息的手段和途径。目前，国内外已建立了大量的平行语料库，但是除了小规模、有限领域的平行语料库外，作为基于汉维两种语言的，通用、较大规模的汉维平行语料库还没有得到开发和建立，汉维双语句子对齐问题也还没有完全解决。

本文以建立一个通用的句子级对齐的汉维平行语料库为基本出发点，围绕与平行语料库的构建密切相关的语料库设计、语料收集、标注加工、双语句子对齐以及语料库管理和应用等问题进行研究。

本文的主要研究内容有：

1）平行语料库的设计规划

语料库的总体框架和设计直接影响到语料库的研究和应用价值。在这方面，本文首先讨论了语料库规划的目的、方法、内容等问题。然后重点介绍了汉维平行语料库的建设目标、基本思路以及工作步骤等总体设计。

2）语料的收集

本文研究的语料库是通用的汉维平行语料库，根据一库多用的建库目标和汉维语的语言特点，结合建库过程中所遇到的问题，讨论了语料收录原则与要求、语料采集方法与来源、语料的筛选与整理、代表性以及语料库的版权等与语料的收集有关的问题。介绍了汉维平行语料库的规模和各种语料的比例以及语料采集预处理工具。

3）平行语料库的标注

收集语料之后的关键一步就是利用某种符号系统把预期有用的信息在语料库中加以标注。在汉维双语语料标注方面现有研究成果不多的情况下，本文在介绍语料标注的目的、规范、编码语言、标注的范围等一般问题的基础上，重点讨论了汉维平行语料库的标注和加工问题。

汉维平行语料标注时，尽可能采纳国际通用的标注标准，首先制定了基于 XML 语言的标注规范和标记集。然后运用标注工具进行了文本属性信息、结构信息和语法信息的标注。我们认为，基于 XML 的标注格式便于资源共享、统一处理和交换，既符合国际标准，又符合汉维平行语料库的要求。

4）维语的词法标注问题

在汉维平行语料库中，我们需要对维语文本进行词法层面的分析与标注。

目前却没有一个共享的词法分析软件可以运用，大部分软件还处于研究开发状态。本文将维语的词法分析当做一个重点内容，对逆向最大匹配方法进行了较深入的探讨，并在此基础上对维语电子词典的结构和搜索算法做了部分改进，在词典中加进去了词干的概率信息。设计并开发了基于反序词典的维文词性标注系统。结果表明，该算法成功解决了语音弱化等关键问题，加快了维文的词语切分和词性标注的速度，使得系统的效率有了显著提高。

5）汉维双语句子对齐

汉维双语句子对齐是本文研究的重点之一。本文在吸取不同句子对齐方法优点的基础上，将“锚点”的概念提升到句子层面，提出了基于锚点句的分段句子对齐算法。该方法首先采用词汇信息和长度信息相结合的方法，找出满足一定条件的锚点句对，然后以锚点句对作为分割标志对全文进行分段和句子对齐。由于充分利用了词汇、数字、标点符号信息，该方法的领域移植性好。由于采用了“分段”方法，既避免了复杂的计算过程，又有效解决了错误蔓延问题。最后的实验结果也表明，该方法的效率较高，在多领域文本中汉维句子对齐的正确率达到了 94.6%。

6）语料库应用工具的开发和应用

一个平行语料库建成以后，是否能得到充分的应用，关键在于该语料库是否具备能满足使用者要求的检索和统计功能。本研究所开发的检索（应用）工具实现在双语语料库中对齐语句的自动链接式检索。该工具具有查询、检索、统计、打印等功能，支持汉文和维文的模糊检索和复杂检索，实行双语语料库的词频统计和检索结果自动排序。该工具所提供的功能在语料库的应用方面起到重要作用。最后指出了汉维平行语料库系统的使用前景和潜在价值。

总之，本文结合汉维平行语料库建设的实践，深入讨论了平行语料库的设计、语料的采集、标注、句子对齐、语料库的检索与应用等问题。

关键词：平行语料库 汉语 维语 维文词法分析 句子对齐 锚点句对

ABSTRACT

Parallel corpus is a corpus that consists of source texts in one language and their translated texts in another language. The main objectives of parallel corpus construction are collecting and annotating large number of machine readable bilingual translated texts in order to find out the translation equivalents of different languages from the parallel corpus by using indexing tools, and to provide researches in linguistics and other fields with the resources and methods. Until now, many parallel corpora have been built in both at home and abroad, however, there is not a large scale, general Chinese-Uyghur parallel corpus except some small and special domain parallel corpora. Many problems, for instance sentence alignment, have not been solved yet.

In this paper, parallel corpus related issues such as corpus design, text collection, corpus annotation, bilingual sentence alignment, corpus management and application, are studied on the base of constructing a sentence aligned Chinese-Uyghur parallel corpus. The main results of the paper are as follows:

1) Parallel corpus design

The design and main structure of a corpus have a big influence on its future application and value. Purposes, methods and aspects of corpus design are analyzed firstly. Then, whole design, including the goal, methodology, and process, of Chinese-Uyghur parallel corpus is introduced.

2) Text collection

The Chinese-Uyghur parallel corpus under construction is a general corpus to several purposes. During the corpus text collection we have encountered with many problems that we have to consider. This paper discusses such problems as principles, requirements, resources, methods, cleaning up process, representativeness, copyright issue during the text collection, and also introduces the scale of the Chinese-Uyghur parallel corpus, and tools used for text collection.

3) Parallel corpus annotation

Another important step after text collection is corpus annotation. In the case of very few works related to Chinese-Uyghur parallel corpus annotation, this paper discusses the purpose, standard, tagset, metalanguage and scope of annotation, and also introduced the annotation process of Chinese-Uyghur parallel corpus.

During the annotation, at the first step, we established a standard of corpus annotation on the base of XML annotation standard. Secondly, we annotated the attributes and structure of all texts, and their grammatical information by annotation tools. We found that XML annotation is a true option for us because of its high

efficiency and advantages during information exchange.

4) Uyghur morphological analyses

In the Chinese-Uyghur parallel corpus, we have to analyze and annotate the Uyghur text at word level. However, there is not any available tool or software for this purpose, though some tools are under study. So, this paper also studies the Uyghur morphological analyses as its important part. We used RMM (Reverse Directional Maximum Matching) algorithm for lemmatization of Uyghur words, improved the whole structure of dictionary, and also added the probability information to each words in dictionary. On this base, we developed the software of Uyghur morphological analyses. The result shows that the method can solve problem of weakening of vowels, and it is highly efficient.

5) Sentence alignment

Chinese-Uyghur sentence alignment is an important part of this study. The paper, on the base of other sentence alignment methods, introduced anchor sentence based two step sentence alignment method. In the first step, we used some lexical and length information to generate anchor sentences that satisfy the condition. In the second step, texts are divided into several sections by using anchor sentence as boundary, and then sentences in each section are aligned. This method avoids the complex computing and error spreading because of its "subsection" technique. Experiment result shows that precision of the method is 94.6% on the average for Chinese-Uyghur sentence alignment in multi-domain texts.

6) Development and application of corpus tools

After building a corpus, it is very important to have a corpus application tool with several functions such as indexing, searching, statistic, without this we can not use the corpus. So, we developed the parallel corpus application software, which works with Chinese-Uyghur parallel corpus. It can index KWIC lines of both Chinese and Uyghur texts at the same time, sort and output search results in different formats. This tool will be an important facility of Chinese-Uyghur parallel corpus use. At last, the paper also points out potential value and application of the Chinese-Uyghur parallel corpus.

In a word, this paper discusses the design, text collection, annotation, sentence alignment, application of parallel corpus through the Chinese-Uyghur parallel corpus construction practice.

Keywords: Parallel corpus; Chinese; Uyghur; Uyghur morphological analyze; Sentence alignment; Anchor sentence

目　录

图目录

表目录