国家出版基金项目
NATIONAL PUBLICATION FOUNDATION

新闻出版改革发展项目库入库项目

“十二五”国家重点图书

·藏文信息处理技术·

བོད་སྐད་སྐད་ཆས་རིག་པ།

藏语计算语言学

欧 珠　扎西加　编著

SWJTUP 西南交通大学出版社

国家出版基金项目
NATIONAL PUBLICATION FOUNDATION

新闻出版改革发展项目库入库项目

“十二五”国家重点图书

西藏大学博士学位授权立项建设项目

藏文信息处理技术

 བོད་སྐད་སྐད་ཆོས་རིག་པ།

藏语计算语言学

欧 珠　扎西加　编著

西南交通大学出版社
·成　都·

内容简介

本书是国内第一本全面系统地介绍藏语计算语言学的著书，由浅入深地讲解了藏语计算语言学的基本理论和知识框架。全书在汉语、英语等自然语言计算机处理的相关基本知识和方法的基础上，结合藏语特点，介绍了基于规则的藏语自然语言分析方法，也介绍了基于统计的分析方法。全书共分四个部分，分别为藏文及藏文信息处理的基础知识、藏文词法分析、藏文句法分析和藏文语义分析。第一部分介绍了计算语言学的基本理论知识、藏文及藏文信息处理的基本知识；第二部分针对藏文信息处理中特有的分词问题，介绍了藏文自动分词及分词规范、分词中歧义的消解、未登录词识别、藏文词性标注及标注标准，以及藏文语料库的相关知识；第三部分从藏语语法的表示入手，将藏语自然语言处理形式化，再给出藏语语法分析的算法；第四部分介绍了藏语语义的表示及分析算法。

本书可作为高等学校和科研院所计算机专业高年级本科生、藏文信息处理方向研究生的教材，也可以作为从事藏语自然语言处理应用领域的研究人员和技术人员的参考资料。

图书在版编目（CIP）数据

藏语计算语言学 / 欧珠，扎西加编著. —成都：西南交通大学出版社，2014.11
（藏文信息处理技术）
ISBN 978-7-5643-2206-9

Ⅰ. ①藏… Ⅱ. ①欧… ②扎… Ⅲ. ①藏语－计算语言学 Ⅳ. ①H214

中国版本图书馆 CIP 数据核字（2013）第 035109 号

藏文信息处理技术
藏语计算语言学
Zangyu Jisuan Yuyanxue
欧 珠 扎西加 编著
*
责任编辑 张 雪 黄庆斌 宋彦博
封面设计 墨创文化
西南交通大学出版社出版发行
成都市金牛区交大路 146 号 邮政编码：610031 发行部电话：028-87600564
http: //www.xnjdcbs.com
四川森林印务有限责任公司印刷
*
成品尺寸：210 mm × 285 mm 印张：15.25
字数：381 千字
2014 年 11 月第 1 版 2014 年 11 月第 1 次印刷
ISBN 978-7-5643-2206-9
定价：38.00 元

前　言

计算语言学（Computational Linguistics）是一门集语言学、计算机科学、数学与逻辑学、认知科学等多种学科于一身的交叉性学科。它以自然语言为处理对象，以计算机科学为研究工具，以数学为自然语言的建模工具，旨在通过建立形式化的数学模型来分析、处理自然语言。其内容涉及语言学、信息科学、人工智能、认知科学、数学（数理逻辑、自动机和形式语言、图论、统计学等）等多个领域。例如，用计算机对自然语言的音、形、义等信息进行处理即对字、词、句、篇章进行输入、识别、分析、理解、生成等。

纵观藏文信息处理技术 20 多年来的发展，在国家强有力的支持下，我国藏语文也跟其他语言文字一样，大步跨上了信息高速公路。如今藏文字符有了自己的国际标准和国家标准，众多科研人员不仅研发出了藏文信息处理的基础和共性软件，实现了藏文字的信息处理，而且在藏文信息处理的知识获取和应用开发方面进行了不懈努力和有益探索，取得了不少成绩，为开展藏语计算语言学研究奠定了坚实的基础。藏语计算语言学学科形成时间虽短，但发展速度很快，是在我国长期从事藏语语法研究基础上开展的藏文信息技术研究，是积极探索学科增长点的结果，是多学科优势互补、相互融合的结晶。

无论是在校计算机和藏文信息技术专业的学生，还是工作在藏语自然语言处理领域的工程技术人才，都需要夯实自然语言处理的基础。为此，我们根据近年来承担的国家有关项目，参考了大量计算语言学方面的书籍，在汉语自然语言处理方法的基础上，根据藏语特点和用户的需求，第一次系统地介绍了藏语计算语言学的基本理论和知识框架，内容包括藏语自然语言处理中词法、语法和语义处理的理论与技术，并结合相关的科研项目，利用其科研成果，实现了自然语言处理的藏语本地化。

作为中文信息处理研究的一个分支，藏语计算语言学所进行的研究在中文信息处理的范畴之内。这门学科的出现，是中文信息处理研究的纵深发展，也是中文信息处理研究领域的拓展。在计算语言学这门学科内容中引入藏语知识描述与处理，无疑给我国计算语言学学科的发展注入了新鲜的血液，从不同语言和思维模式中探索新的方法与技术，可使之日臻完善。

本书是在西藏大学博士学位授权学科“藏语计算语言学”立项建设项目、藏文信息技术教育部创新团队（编号：IRT0975，名称：藏文信息技术“长江学者和创新团队发展计划”创新团队）、国家自然科学基金项目“藏语依存树库的构建”（批准号：61163043）、“计算机及藏文信息技术”国家级教学团队、西藏大学珠峰学者等项目的资助下所完成的项目成果之一。

本书由西藏大学藏文信息技术国家地方联合工程中心策划，由欧珠、扎西加编著。本书

在编写过程中得到了北京大学计算语言学研究所俞士汶教授、王厚锋教授、常宝宝副教授、詹卫东副教授，中国科学院计算技术研究所宗成庆研究员，中国标准化研究院基础标准化研究所陈玉忠研究员，西北民族大学多拉教授，西藏大学藏文信息技术国家地方联合工程研究中心大罗桑朗杰教授等专家的关键性指导和支持。特别是多拉教授和常宝宝副教授审阅了本书的初稿，并提出了许多宝贵的意见。我们谨在此一并表示衷心的感谢。

本书作为“藏文信息处理技术”丛书之一，被列入 2013 年度新闻出版改革发展项目库入库项目（项目序号：OO20130845）、《“十二五”国家重点图书、音像、电子出版物出版规划》项目（新广出办发〔2014〕75 号-294）、国家出版基金项目（项目编号：2014T_2-001）。

如果本书对读者有所帮助，我们将感到万分荣幸。由于作者水平所限，书中难免存在不妥之处，恳请广大读者批评指正。

作　者

2013 年 1 月

目　录

第 1 章

计算语言学概论

1.1 计算语言学的定义

计算语言学（Computational Linguistics）是一门集语言学、计算机科学、数学与逻辑学、认知科学等多种学科于一身的交叉性学科。它以自然语言为处理对象，以计算机科学为研究工具，以数学为自然语言的建模工具，旨在通过建立形式化的数学模型来分析、处理自然语言。其内容涉及语言学、信息科学、人工智能、认知科学、数学（数理逻辑、自动机和形式语言、图论、统计学等）等多个领域。例如，用计算机对自然语言的音、形、义等信息进行处理即对字、词、句、篇章进行输入、识别、分析、理解、生成等。

计算语言学是伴随着计算机的诞生而产生的，最早是由应用性研究促发而来的。在 1946 年第一台计算机诞生后不久，就有人想到用它来进行语言之间的翻译，这就是机器翻译的起源，也是计算语言学的发端。机器翻译是计算机与自然语言的第一个结合点。1966 年在美国科学院语言自动处理咨询委员会（Automatic Language Processing Advisory Committee，ALPAC）的一份对机器翻译译文质量的评估报告中首次提出了“计算语言学”的概念。

英国《大不列颠百科全书》中对计算语言学的定义：计算语言学是利用电子数字计算机进行的语言分析。虽然许多其他类型的语言分析也可以运用计算机，但计算机分析最常用于处理基本的语言数据，例如建立语音、词、词元素的搭配以及统计它们的频率。

《现代语言学词典》（戴维·克里斯特尔，1997）中对计算语言学的定义为：语言学的一个分支，用计算技术和概念来阐述语言学和语音学问题。已开发的领域包括自然语言处理、言语合成、言语识别、自动翻译、编制词语索引、语法的检测，以及许多需要统计分析的领域。

语言学家刘涌泉在我国的《大百科全书》（2002）中的解释是：计算语言学是语言学的一个分支，专指利用电子计算机进行语言学习。

综合以上定义，计算语言学实际上包括以语音为主要研究对象的语音学基础及其语音处理技术研究，以及以词汇、句子、话语或语篇及其词法、句法、语义和语用等相关信息为主要研究对象的处理理论与技术研究。

1.2 计算语言学的发展历程

在“计算语言学”这个术语出现之前，就有一些具有远见卓识的学者研究过语言的计算

问题，他们从计算的角度来研究语言现象，揭示语言的数学面貌。关于语言计算的思想和研究是源远流长的，其中有四项基础性的研究特别值得注意：

（1）马尔可夫（A.Markov）关于马尔可夫模型的研究；

（2）图灵（A.M.Turing）关于算法计算模型的研究；

（3）香农（C.E.Shannon）关于概率和信息论模型的研究；

（4）乔姆斯基（N.Chomsky）关于形式语言理论的研究。

早在 1913 年，俄罗斯著名数学家马尔可夫就注意到俄罗斯诗人普希金的叙事长诗《欧根·奥涅金》中语言符号出现概率之间的相互影响。他试图以语言符号的出现概率为实例，来研究随机过程的数学理论，由此提出了“马尔可夫链”（Markov chain）思想。他的这个开创性的成果用法文发表在俄罗斯皇家科学院的通报上。后来马尔可夫的这一思想发展成为在计算语言学中广泛使用的马尔可夫模型（Markov model），是当代计算语言学最重要的理论支柱之一。

在计算机出现以前，英国数学家图灵就预见到未来的计算机将会对自然语言研究提出新的问题。1936 年，图灵向伦敦权威的数学杂志投了一篇论文，题为《论可计算数及其在判定问题中的应用》。在这篇开创性的论文中，图灵给“可计算性”下了一个严格的数学定义，并提出了著名的“图灵机”（Turing machine）数学模型。“图灵机”不是一种具体的机器，而是一种抽象的数学模型，据此可制造一种十分简单但运算能力极强的计算装置，用来计算所有能想象得到的可计算函数。1950 年 10 月，图灵在《机器能思维吗？》一文中指出：“我们可以期待，总有一天机器会同人在一切的智能领域里竞争起来。但是，以哪一点作为竞争的出发点呢？这是一个很难决定的问题。许多人以为可以把下棋之类的极为抽象的活动作为最好的出发点，不过，我更倾向于支持另一种主张，这种主张认为，最好的出发点是制造出一种具有智能的、可用钱买到的机器，然后，教这种机器理解英语并且说英语。这个过程可以仿效小孩子说话的那种办法来进行。”①

图灵提出，检验计算机智能高低的最好办法是让计算机来讲英语和理解英语，进行“图灵测试”。他天才地预见到计算机和自然语言将会结下不解之缘。

20 世纪 50 年代提出的自动机理论来源于图灵在 1936 年提出的可计算性理论和图灵机模型。图灵划时代的研究工作被认为是现代计算机科学的基础。图灵的工作首先催生了麦库洛克·皮特（McCulloch-Pitts）的神经元（neuron）理论，即一个简单的神经元模型就是一个计算的单元，它可以用命题逻辑来描述。其次，图灵的工作还引导了克林（Kleene）关于有限自动机和正则表达式的研究。

1948 年，美国学者香农使用离散马尔可夫过程的概率模型来描述语言的自动机。他的另一个贡献是创立了“信息论”（information theory），将通过诸如通信信道或声学语音这样的媒介传输语言的行为比喻为“噪声信道”（noisy channel）或者“解码”（decoding）。香农还借用热力学的术语“熵”（entropy）来作为测量信道的信息能力或者语言的信息量的一种方法，并且用概率技术首次测定了英语的熵。

1956 年，美国语言学家乔姆斯基从香农的工作中吸取了有限状态马尔可夫过程的思想，首先把有限状态自动机作为一种工具来刻画语言的语法，并且把有限状态语言定义为由有限状态语法生成的语言。这些早期的研究工作促使了“形式语言理论”（formal language theory）

① 冯志伟. The Oxford Handbook of Computational Linguistics（牛津计算语言学手册）[M]. 北京：外语教学与研究出版社，牛津：牛津大学出版社，2009.

这样的研究领域的产生，即采用代数和集合论把形式语言定义为符号的序列。乔姆斯基在研究自然语言的时候首先提出了“上下文无关语法”（Context-Free Grammar，CFG）。后来，计算机科学家巴库斯、瑙尔等在描述 ALGOL 程序语言的工作中，分别于 1959 年和 1960 年也独立地发现了这种上下文无关语法。这些研究都把数学、计算机科学与语言学巧妙地结合了起来。

乔姆斯基在计算机出现的初期把计算机程序设计语言与自然语言置于相同的平面上，用统一的观点进行研究和定义。他在《自然语言形式分析导论》一文中，从数学的角度给语言提出了新的定义，指出：“这个定义既适用于自然语言，又适用于逻辑和计算机程序设计理论中的人造语言”。在《语法的形式特性》一文中，他专门用了一节的篇幅来论述程序设计语言，讨论了有关程序设计语言的编译程序问题。这些问题是作为“组成成分结构的语法的形式研究”，从数学的角度提出来，并从计算机科学理论的角度来探讨的。他在《上下文无关语言的代数理论》一文中提出：“我们这里要考虑的是各种生成句子的装置，它们又以各种各样的方式，同自然语言的语法和各种人造语言的语法两者都有着密切的联系。我们将把语言直接地看成在符号的某一有限集合 V 中的符号串的集合，而 V 就叫作该语言的词汇……我们把语法看成是对程序设计语言的详细说明，而把符号串看成是程序。”在这里乔姆斯基把自然语言和程序设计语言放在同一平面上，从数学和计算机科学的角度，用统一的观点来加以考察，对“语言”“词汇”等语言学中的基本概念，获得了高度抽象化的认识。

马尔可夫、图灵、香农和乔姆斯基这四位著名学者对于语言与计算关系的探讨，是早期计算语言学研究的最重要的成果，为计算语言学的理论和技术奠定了坚实基础。

机器翻译是计算语言学最重要的应用领域。1949 年，美国洛克菲勒基金会副主席韦弗（W. Weaver）在一篇以《翻译》为题目的备忘录中，认为翻译类似于解读密码的过程。他说：“当我阅读一篇用汉语写的文章的时候，我可以说，这篇文章实际上是用英语写的，只不过它是用另外一种奇怪的符号编了码而已，当我在阅读时，我是在进行解码。”

早期机器翻译系统的研制受到了韦弗这种思想的很大影响，许多机器翻译研究者都把机器翻译的过程与解读密码的过程相类比，试图通过查询词典的方法来实现词对词的机器翻译，因而译文的可读性很差，难于付诸实用。

由于学者的热心倡导以及业界的大力支持，美国的机器翻译研究一时兴盛起来。1954 年，美国乔治敦大学在国际商用机器公司（IBM 公司）的协同下，用 IBM-701 计算机，进行了世界上第一次机器翻译试验，把几个简单的俄语句子翻译成英语。接着，苏联、英国、日本也进行了机器翻译试验，机器翻译出现热潮。

1952 年，在美国的麻省理工学院（MIT）召开了第一次机器翻译会议。1954 年，出版了第一本机器翻译的杂志，这个杂志的名称就叫作 *Machine Translation*（《机器翻译》）。尽管人们在自然语言的计算方面进行了很多的研究工作，但是直到 20 世纪 60 年代中期，才出现了 computational linguistics（计算语言学）这个术语，而且，在刚开始的时候，这个术语是偷偷摸摸地、羞羞涩涩地出现的。

1965 年，*Machine Translation* 杂志改名为 *Machine Translation and Computational Linguistics*（《机器翻译和计算语言学》）。在杂志的封面上，首次出现了“Computational Linguistics”这样的字眼，但是，“and Computational Linguistics”这三个单词是用特别小号的字母排印的。这说明，当时学者们对于“计算语言学”是否能够算为一门真正的独立的学科还没有把握。计算语言学在刚刚登上学术这个庄严殿堂的时候，还带有“千呼万唤始出来，犹抱琵琶半遮

面”那样的羞涩，以至于学者们不敢用与“Machine Translation”同样大小的字母来排印。当时 *Machine Translation* 杂志之所以改名，是因为在 1962 年美国成立了“机器翻译和计算语言学学会”（Association for Machine Translation and Computational Linguistics），通过改名可以使杂志的名称与学会的名称保持一致。

根据这些史料，我们认为，远在 1962 年，就出现了“计算语言学”这个学科。尽管它在刚出现的时候还是偷偷摸摸的，显示出少女般的羞涩，但是最后，计算语言学这个新兴的学科终于萌芽了，它破土而出，悄悄地登上了学术的殿堂。

经过将近 60 年的发展，计算语言学的研究出现了空前繁荣的局面。这主要表现在如下三个方面：

第一，概率和数据驱动的方法几乎成为了计算语言学的标准方法。在句法剖析、词类标注、参照消解、话语处理、机器翻译的算法中都开始引入概率，并且采用从语音识别和信息检索中借过来的基于概率和数据驱动的评测方法。

第二，计算机速度的加快和存储量的增大，使得在计算语言学的一些应用领域，特别是在语音合成、语音识别、文字识别、拼写检查、语法检查这些应用领域，已经进行了卓有成效的商品化开发。自然语言处理的算法开始被应用于“增强交替通信”中，语音合成、语音识别和文字识别的技术也已经应用于“移动通信”中。

第三，随着网络技术的发展，互联网逐渐变成一个多语言的网络世界。互联网上多语言的机器翻译、跨语言信息检索正在迅猛发展，计算语言学的各种应用技术事实上已经成为互联网技术的重要支柱。

1.3 计算语言学的研究范畴

计算语言学的研究内容十分广泛，根据其应用领域的不同，大致可以分为以下几个方面：

（1）机器翻译（machine translation）：实现从一种语言到另一种语言的自动翻译。

（2）文语转换（text to speech）/语音合成（speech synthesis）：计算机自动地把给定的文本信息转换成语音的过程。

（3）语音识别（speech recognition）：计算机通过识别和理解过程把语音信号转变为相应的文本或命令。

（4）文本分类（text categorization）：计算机对文本集（或其他实体或物件）按照一定的分类体系或标准进行自动分类标记。

（5）信息检索（information retrieval）：或称情报检索，就是利用计算机从海量文档中找到符合用户需要的相关文档。

（6）信息抽取（information extraction）：从海量的信息中抽取出用户所需要的（结构化）信息。

（7）文字校对（text-proofing）：对文字的拼写、用词，甚至语法、文档格式等进行自动检查、校对和编排。

（8）问答系统（question answering）：信息检索系统的一种高级形式，它能用准确、简洁的自然语言回答用户用自然语言提出的问题。

（9）自动摘要（automatic summarization）：将原文档的主要内容或某方面的信息自动提取出来，并形成原文档的摘要或缩写。

（10）语言教学（language teaching）：借助计算机辅助教学工具，进行语言教学、操练和辅导等。

实际上，在语言学的很多领域需要研究解决的问题，都会是计算语言学的研究内容。

1.4　计算语言学的研究方法

在计算语言学的发展过程中，提出了很多方法。这些方法，在理论上有一定的深度，在实践上也有实用价值。对于计算语言学方法的研究，可以从方法论的角度来论述，也可以从语音、词汇、形态、句法、语义、语用研究中使用的方法来论述。从方法论的角度，计算语言学方法可以分为基于规则的方法（rule-based approach）和基于统计的方法（statistics-based approach）两个方面。基于规则的方法是理性主义（rationalism）的方法，基于统计的方法是经验主义（empiricism）的方法。这两种方法实际上并不是完全对立的，它们各有利弊，而且目前这两种方法有合流的倾向，它们正在相互结合起来，取长补短，相得益彰。

多年来，在计算语言学和自然语言处理的研究中，理解和生成语句的必经之路是句法分析和语义分析。因此在很长一段时间里，许多语言处理系统都是基于规则的。建立这样的系统，从整体构架到具体的处理技术，语言学的研究都是必不可少的基础。基于规则的方法最早应用在机器翻译当中，被证明在一定范围内是相当有效的。随着研究的深入和应用目标的发展，人们逐渐发现，实际上很难用规则的形式把各种语言事实和理解语言所需要的背景知识充分地表达出来。因此，基于规则的自然语言处理系统往往只能在极其受限的某些语言环境中获得一定的成功。

经验主义者认为基于规则的方法存在以下缺陷：一是通过内省方式得到的语言规则往往有脱离言语实际的可能；二是规则的灵活性较差，易忽略语言中那些经验性的、小粒度的知识，难以覆盖各种复杂纷繁的语言现象；三是当需要添加新的规则时，又必须注意协调与已有规则的关系，避免规则之间产生矛盾。为了克服这些局限，他们用统计学方法从大规模语料中分析和归纳语言现象和规律，再用得到的统计规律或语言模型来处理自然语言。他们认为这些统计规律或语言模型体现了从真实语料中直接获取的语言知识，不但可以用于语言信息处理系统，而且能用来检验语言学研究中依靠手工搜集材料的方法所得出的结论。基于语料库的方法自 20 世纪 90 年代初从国外引入以后，在自然语言处理的许多分支领域得到了应用。其中最有成效的当数语音识别与合成，在书面语处理的各个层面上（词语、句法、语义），这种方法几乎都有用武之地。其中有代表性的应用是汉语自动分词和词性标注，文本信息检索，信息抽取，信息过滤和文本自动分类。与基于语料库方法有关的研究工作包括：各种语言成分的对齐（alignment），语言知识的学习和获取，各种语言统计模型的建立，语言成分相似度的计算等。语料库方法的优点是可以使语言现象数量化，这非常适合计算。目前的研究大致可以分为两类：一是基于简单相关统计的方法，也可以称为语言资源性分析；二是在统计意义下的建模方法，这涉及机器学习的算法问题。经过几年的研究和实践，人们也开始

对统计语言模型本身进行分析和再认识，提出了适用的统计模型及其适用范围、统计量的繁与简、统计对象的升华、多统计量的结合等问题。

基于语料库的方法当然也有自己的局限性。在人们对语言的机制还缺乏系统了解，还没有一种适合信息处理的语言理论可以应用的时候，经验主义实际上是一种依靠“量”来获取“质”的策略。也就是说，大量语言现象的统计规律能够确切地反映语言的结构规律和言语过程的认知规律，而实际上这还是有待证明的观点。我们还不能从语言学的角度解释所采用的统计方法和语言模型，说明统计数据的语言学意义。统计语言模型需要建立在语言学知识的基础上，一个语言模型能否达到比较好的处理效果，很大程度上取决于我们能为它提供什么样的语言学知识作为参数。目前统计语言模型在机器翻译系统中效果不佳，其主要原因应该不是模型本身的计算能力有限，而是能够提供给模型的关于机器翻译的知识太少。我们目前的基础研究还没有发掘出足够的语言学知识，或者是还没有把这些知识系统地、结构化地组织起来。这一点对基于规则的方法来说，同样也是亟待解决的问题。

把两种方法结合起来，取长补短，互为补充，也许是更好的办法。目前已有不少语言信息处理系统采用了这种混合策略。譬如在机器翻译中用规则分析句子的句法结构（有些语言现象用规则处理比较方便），用基于语料库的方法处理词语的搭配问题。经验主义者也认识到，即使算法的主要基调是基于统计形式的（例如向量空间模型），规则形式仍然能够找到合适的位置，例如那些概率接近 1 的规则，即几乎无例外的规则。

1.5 计算语言学与藏语研究

计算语言学是以语言学为研究基础，其内容涉及计算机科学、逻辑学和心理学等多种学科的一门典型的交叉学科。语言是思维的载体，是人际交流的重要工具。在人类历史上以语言文字形式记载和流传的知识占到知识总量的 80%以上，就计算机的应用而言，85%左右是用于语言文字的信息处理。在这样的社会需求下，计算语言学作为语言信息处理技术的一个重要方向，在推动人类文明的进程中发挥着重要的作用。

藏文作为有着 1 300 多年历史的古老文字，承载了藏民族光辉灿烂的文化，并表达着独具特色的民族意识和思维方式。用藏文记载的经典文献、古籍著述和译作浩如烟海，在我国境内仅次于汉文文献。然而，“自从人类进入以计算机和网络为主体的信息时代，古老的藏文字正面临着一场‘生死存亡’的考验——即能否跨入信息时代。藏文字一旦不能跨入信息时代，它必将失去语言文化载体的基本功能和作用，就会被这个时代无情地抛弃。纵观国内外语言文字信息处理技术的发展历史和现状，我们可以清楚地看到，古老的藏文字能否跨入信息时代的关键就是能不能解决好藏文信息处理技术问题。”[①]因此，藏文信息处理是直接关系着藏文命运的一件大事，其重要意义是不言而喻的。

计算语言学作为文本信息处理的一般原理与方法、技术，越来越受到重视并正在成为一个具有战略性特点的学科。随着我国综合国力的增强，计算技术水平的提高和互联网应用的迅猛发展，计算语言学与藏文信息处理迎来了一个非常好的发展环境。藏文信息处理的研究

① 陈玉忠，俞士汶. 藏文信息处理技术的研究现状与展望[J]. 中国藏学，2003（64）：97-107.