

中国民族语言语法标注文本丛书

白语语法标注文本

王锋 / 著

江荻 / 主编

本套语源建设丛书具有典型真实文本和
跨语言本对号特征，是语言类型学、历史语言学、区域语言学、
语言接触研究等资源库模型等不可或缺的语言基础和
学术财富，从事语言学、民族学、人类学、
历史学和自然语言处理等学科研究的资源库。



国家出版基金项目
NATIONAL PUBLISHING FUND PROJECT



社会科学文献出版社
SOCIAL SCIENCES PRESS (CHINA)

中国民族语言语法标注文本丛书

江荻 / 主编

 社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS (CHINA)

白语

语法标注文本

王锋 / 著



国家出版基金项目
NATIONAL PUBLISHING FUND PROJECT

图书在版编目(CIP)数据

白语语法标注文本 / 王锋著. -- 北京: 社会科学
文献出版社, 2016. 11

(中国民族语言语法标注文本丛书 / 江荻主编)

ISBN 978 - 7 - 5097 - 9582 - 8

I. ①白… II. ①王… III. ①白语 - 语法 - 研究

IV. ①H252. 4

中国版本图书馆 CIP 数据核字 (2016) 第 194336 号

· 中国民族语言语法标注文本丛书 ·

白语语法标注文本

主 编 / 江 荻

著 者 / 王 锋

出 版 人 / 谢寿光

项目统筹 / 宋月华 周志静

责任编辑 / 孙以年

出 版 / 社会科学文献出版社·人文分社 (010) 59367215

地址: 北京市北三环中路甲 29 号院华龙大厦 邮编: 100029

网址: [www. ssap. com. cn](http://www.ssap.com.cn)

发 行 / 市场营销中心 (010) 59367081 59367018

印 装 / 三河市尚艺印装有限公司

规 格 / 开 本: 787mm × 1092mm 1/16

印 张: 18 字 数: 418 千字

版 次 / 2016 年 11 月第 1 版 2016 年 11 月第 1 次印刷

书 号 / ISBN 978 - 7 - 5097 - 9582 - 8

定 价 / 128.00 元

本书如有印装质量问题, 请与读者服务中心 (010 - 59367028) 联系

 版权所有 翻印必究

《中国民族语言语法标注文本丛书》编辑委员会

主 编 江 荻

副主编 黄 行 燕海雄

编 委

白碧波 陈国庆 黄 行 江 荻 蓝利国 李大勤

林幼菁 刘 宾 刘 岩 龙从军 鲁美艳 马 辉

苗东霞 潘立慧 千玉花 斯钦朝克图 王 锋

王海波 韦学纯 徐世璇 许鲜明 燕海雄 杨将领

杨 艳 尹巧云 尹蔚彬 张 军 钟耀萍

基金资助项目:

中国社会科学院语音学与计算语言学重点实验室 (2009~2011):
中国民族语言文本标注研究

中国社会科学院重大课题 (2011~2013):
中国民族语言语法标注文本丛书 (YZDA2011-18)

国家社科基金重大招标项目 (2011~2016):
中国民族语言语法标注文本及软件平台 (10&ZD124)

中国社会科学院创新工程 (2013~2015):
中国民族语言语料工程及深度应用研究

国家社科基金重大招标项目 (2013~2018):
基于大型词汇语音数据库的汉藏历史比较语言学研究 (12&ZD174)

国家自然科学基金重点项目 (2012~2016):
基于本体的多策略民汉机器翻译研究 (61132009: 藏汉机器翻译)

前 言

在中国民族语言研究历程中，资源和语料建设一直是重中之重。语料的形式和内容多种多样，譬如词汇、词典、文本、音档、语图、语音参数、文字图片、多语对照词汇、语言或方言地图，以及人名、地名等其他各类专题语料资源。

通过图书出版而面世的语料主要有各种民族语言描写专著提供的案例，特别是其中附载的词汇和文本，这是所谓单一语言或方言语料的常见汇集形式。零星出版的这类专著很多，此处不能一一列出，而以丛书形式发布的则影响较大，主要有“中国少数民族语言简志”丛书（近 60 卷）、“中国新发现语言研究”丛书（40 余卷）和“中国少数民族语言方言研究”丛书（近 20 卷），以及“中国少数民族语言系列词典”丛书（20 余卷）。此外，近年一批以“参考语法”为题的博士学位论文大多也附带一定数量的分类词汇和篇章文本。至于涉及多种语言或方言语料的各语族论著也蔚为大观，例如孙宏开主编的《藏缅语语音和词汇》、黄布凡主编的《藏缅语族语言词汇》、王辅世和毛宗武合著的《苗瑶语古音构拟》、梁敏和张均如合著的《侗台语族概论》、严其香和周植志合著的《中国孟高棉语族语言与南亚语系》（7 种语言 14 个方言点）、孙竹主编的《蒙古语族语言词典》（6 种语言 16 个方言点）、陈宗振主编的《中国突厥语族语言词汇集》（8

种语言)、朝克编著的《满通古斯语族语言词汇比较》(6种语言),等等。

随着信息化时代的发展,21世纪以来,前期调查和出版的相当部分词汇数据进入了电子化资源检索平台,改变了语言学家的工作方式和 workflows,拓宽了他们的研究领域和研究方向,增强了他们驾驭语言资源的能力,甚至推动他们创造出新的语言学说和方法。据我们了解,这些电子化数据资源中影响较大的有“汉藏语同源词检索系统”和“东亚语言词汇语音数据检索系统”。有研究表明,这两个系统为学术研究的深度发展提供了新的契机,解决了不少研究中的疑难问题。

可是,以上所述成果形式无论是附着于描写或专题论著还是独立资源著作,似乎主要集中在各类民族语言的词汇表和词典方面,说明学界历年积累的资源还有重大空白,尤其缺乏文本性质的熟语料标注资源。

随着语言研究的深入和研究领域的拓展,特别是伴随着语言类型学(语法类型、语音类型等)、普遍语法、语系学说、语言接触、语言心理、语言生态、语言检索和多语机器翻译等新兴跨学科研究在中国的蓬勃兴起,学术界开始呼唤一种跨语言、跨方言的资源性建设研究,呼唤创造多样性的真实文本资源和跨语言对齐文本资源。值得称道的是,中央民族大学少数民族语言文学学院适时推出了一套“中国少数民族语言话语材料”丛书,迄今已出版黎语、临高语、佯语、仡佬语、布央语、布依语、撒拉语等双语对照文本材料,初步弥补了该领域的不足。

约20年前,北京大学老朋友郭锐教授跟我聊起民族语言,询问我民族语言研究领域是否有文本性语篇材料。我当时一愣,老实回答他尚无此类资源。其时,我刚刚主持完成“中国少数民族语言研究文摘数据库系统”,显见,当时的思路还处在仅仅为研究者提供研究信息的阶段。1998年,孙宏开先生和丁邦新先生合作开展汉藏语同源词研究,我受命研制电子版同源词检索系统。此后进入21世纪,我又承担了研制东亚语言词汇语音检索系统的工作。也许是学术使命使然,我并没有忘记郭锐之问,开始在民族语言学界推动文本资源开发。最初我将世界少数民族语文研究院(SIL)的文本处理工具 toolbox 教学软件资料编译成中文,2006年起在多所高校讲授。2009年,我们实验

室举办 toolbox 培训班,跟部分民族语言专家签署开发标注语料协议。2010 年我们得到中国社会科学院重大课题(YZDA2011-18)支持,这就走上了“中国民族语言语法标注文本丛书”的研制道路,其后又进一步得到国家社科基金重大项目(10&ZD124)的支持和中国社会科学院创新工程项目“中国民族语言语料工程及深度应用研究”的支持。这是本丛书研制的基本背景。

这套丛书有多方面的价值和特征。

(1) 创新范式。在描写语言学领域内,以往传统观念总是把记录语料作为语法著作的附录,数量少且处于附属地位。这套丛书虽然也安排了语言概况或语法导论,却以服务于作为正文的标注文本为目的,这种以传统著作附录作为正篇的研制思路既是对文本语料缺乏的弥补,也开拓了语言研究的新方向。这跟学界倡导的记录语言学不谋而合。更具价值的是,丛书作者所采纳的文本大多来自田野调查,或来自民间记录故事。与以往的例句翻译式调查或诱导式例句调查相比,这样的语料从本源上避免了主观性,甚至杜绝了母语人自身的内省式语法案例。从方法论上看,以真实文本为语料的研究很可能引起中国描写语言研究范式的变革,这未尝不是好事。

(2) 基础标注。课题组提出一个关于标注的基本标准,即描写语法的基础标注。这么做是基于我们为语言专题深度研究提供支撑的服务理念,包括服务于作者自己的深度专题研究。我们从三方面加以说明。其一,我们认为新近发展的一些语言分支学科具有资源依赖性质,例如语言类型学一般是跨语言或跨方言的,语言接触研究也需要双语或多语言资源的支持。对于无文字的语言,它们的语法化或词汇化研究更需要亲属语言的相互印证。至于机器翻译也一定是在双语或多语语料相互对照条件下才能开展起来的。其二,丛书中有藏缅语言、侗台语言、苗瑶语言、南亚语言,还有阿尔泰语言,语言自身类型差异很大,譬如一些语言是 SVO 语序,另一些则是 SOV 语序,有些是前置词系统,有些则是后置词(词格)系统,等等。特别是目前各语言研究的广度和深度差异较大,采纳的理论和研究的方法也不完全相同,为此,确定一个简洁的基本结构方法或描写方法对文本进行语法标注是合适的。其三,业有所长,术有专攻。真正利用这套

丛书语料的学者未必熟悉各种语言，更不可能很快掌握这些陌生语言的语法体系，要求每个学者都调查多种语言、掌握多种语言并不现实，也没必要。在这个意义上，我们组织合适的专业人员开发可供其他学者开展专题深入研究的文本资源，特别是熟语料语法标注文本就非常有价值。显然，从以上叙述可以看出，基础标注就是无论某语言是何种类型，无论某语言研究的深度如何，这套丛书都以基本语法结构框架来标注各种语言的词法形态和句法现象，例如“性、数、格、时、体、态”范畴，同时标上通用语对译词语。甚至如果某些语法现象在某种语言中尚未被认识或尚未得到阐释，例如“复指”(anaphora)或“示证”(evidentiality)，则完全可以不标，这样也给使用者留下专题深度挖掘和拓展的空间，这就是描写语法基础标注的意义和价值所在。值得提示的是，这套丛书的作者都是具体语言领域的专家，他们对语言的结构描写和基础标注为读者提供了一个高起点的平台。

(3) 后续可为。中国地广人多，有上百种语言和数千种方言(调查点)，无论从共时还是历时的图景观察，这些多样性的资源都是极为宝贵的人类知识财富。我们组织的第一批文本标注丛书仅仅 12 部，意味着这种研究方法刚刚起步，今后的工作还大有可为。不妨让我们联系美国结构主义的调查方法来看，在这种调查中，所有语言的、文化的和社会的探索都起始于真实文本记录调查，其次是文本标注，包括语法标注；词汇是在标注文本基础上抽取出来的；最终才是文本内容社会的、文化的、人类学的解读。所以我们希望，文本调查和文本标注不仅是一种语言研究的方法，还可以是未来语言研究的一种范式，一种探索文化的范式、一种理解社会的范式。我们期待这套丛书的出版能抛砖引玉，带来更多更好的同类成果。可以说，中国民族语言语法标注资源建设不仅是一种量的积累，而且是一种质的变化，持之以恒，后续工作将创造更有价值的丰富文本资源和学术财富。

为提高丛书可读性，我想对这套丛书的研制方法和阅读要点做一点介绍。

(1) 隔行对照化和标注方式：术语“隔行对照化”来自英语的 Interlinearization，指民族语(大多是小语种)词语跟标注语(通用语，如汉语或英语)以及语法标注的分

行对齐。这种方法是目前世界各国学者研究少数民族语言的主流方法，通过隔行对照化形成一种所有语言学家或语言学专业研究生都能读懂的文本。例如藏语拉萨话：

文字行： ཁོང་ལ་ལོ་ལྔ་ལོ་ལྔ་

原文行： khong lha sar phyin song

分析行： khong lha sa-la vgro -ed song

标注行： 3sg 拉萨-ALA 去-PST ASP-PEF

翻译行： 他去了拉萨。

大多数情况下，“文字行”并不一定需要，无文字语言则无此选项。“原文行”是记录的真实文本话语，多数情况下采用音标记录形式，本例采用了藏文的拉丁转写。“分析行”主要对“原文行”词语加以形态或句法标注，本例 lha sar 书写上包含了向格(la)的文字变体形式(-r)，黏着在前面不带辅音韵尾的音节上；而 phyin 则是动词 vgro(走，去)的过去时形式，很像英语 went 是原形动词 go 的过去时形式，所以“分析行”还原了语法变化前的形式，譬如 phyin = vgro + -ed(过去时等于原形加表示过去时的标记-ed)。“标注行”是对“分析行”进行通语(汉语普通话)标注和语法标注，-ALA 表示向格，-PEF(或 ASP-PEF)表示体范畴的已行体。“翻译行”是原文行的通语直译。

(2)“三行一体”还是两行一体？不少中国民族语言缺乏类似印欧语言的词法形态，即所谓词根语或孤立语。这样一来，“分析行”跟“原文行”基本一致，因此有些语言就不需要采用“三行一体”格式。例如壮语：

原文行： tu³¹ kai³⁵pau⁴² ɕai³⁵ ŋai³¹ tu³¹ ma²⁴ hap³³ tai²⁴

标注行： CL-只 公鸡 又 PASS-挨 CL-条 狗 咬 死

翻译行： 公鸡又被那条狗咬死了。

就我们看到的标注文本，侗台语言、苗瑶语言和部分藏缅语言或许只需原文行和标注行。个别情况下是作者未标出形态变化而无需分析行。

从书中，林幼菁教授撰写的《嘉戎语卓克基话语法标注文本》增加了韵律单位内容，即在展开语法标注之前，先根据口语韵律边界切分文本，然后才标注文本，这样就产生

了韵律分析行。例如：

韵律行：	161	təwamɲepɛ	ʃikoj	
原文行：	təwamɲepɛ		ʃikôj	
分析行：	tə-	wam	=ɲepɛ	ʃikô =j
标注行：	N-	熊	=PL	树上 =LOC

韵律行：	162 ...	təwi	kəzɛpti	ptʂerə
原文行：	təwi	kəzɛpti		ptʂêrə
分析行：	tə-	wi	kə-	za -ɲ =ti ptʂêrə
标注行：	N-	橡实	NMZL-	吃 1 -2/3PL =TOP:OBL 然后
翻译行：	161-162	老熊在树上吃橡实的时候		

我相信，这样的标注为读者提供了更多信息，而且一定会让关注语篇语音现象的专家欣喜。

(3) 标注符号体系。上文拉萨话案例“标注行”中包含了一些语法标注符号，例如 3sg、-ALA、-PST、-PEF 等，这是丛书研制初始建立的语法标注体系。这套标注符号借鉴了国际规范，同时也补充了标注中国语言特定语法现象的符号。为此，课题组建议丛书作者采纳统一的标注符号，但同时也可增加该语言特定需求的符号。所以每一部标注文本著作的前面都列出了标注符号缩写表。

(4) 文本语料的规范与标准。为了实现标注文本的实用性，课题组建议调查或选用的文本具有原生性、连续性、记述性、口传性等特征，而传统口传故事、族群起源传说、儿童或寓言故事、日常口语记录大致包含这些特征，表述通俗、朴实，用词简单、口语化。不过，民间故事口语词汇重复，用词量少，语法结构也过于简单，为了弥补这些不足，也建议选用部分母语作家复杂的民间文学作品，或者少量报刊语体文本；同时，鉴于句类特征（陈述、疑问、祈使、感叹等），还建议选用一两篇问答型对话文本。记录民间故事的时候，发音人是否擅长叙述故事也是很重要的条件。同一个发音人往往风格一致、用词有限，所以尽量选择多个材料提供人和不同题材故事也是较好的策略。课题组还建议书稿作者不选或少选韵文类的诗歌、民歌、唱本之类，这也是为了保证语法现象的完整性和通用性，囊括更多的词汇和语法现象。

(5) 整体布局与对照词汇。每部著作都包含三部分：“1.语法导论”“2.标注文本”“3.对照词汇”。“语法导论”分量不大，主要包括音系、词汇和词法、句法要点。“标注文本”除了句对齐直译，每篇文本之后给出全文翻译。最后的“对照词汇”是从文本中抽取的词汇，即仅列出现于著作文本中的词语，而不是这个语言或方言的任意词汇。词汇基本按照字母顺序排列，部分按照汉语拼音排序。部分著作还列出了词语出现次数。不过，这里需要说明的是，由于排版技术的限制，对照词汇没有列出每个词出现的页码，这算是一件遗憾之事。

这套丛书经历了多阶段和多项课题支持，其中中国社会科学院重大课题和实验室项目于2013年顺利结项，被评定为院级优秀项目，中国社会科学院创新工程项目也于2015年圆满完成。有辛劳就有收获，现在第一批著作即将出版，包括：

《藏语拉萨话语法标注文本》（江荻）

《土家语语法标注文本》（徐世璇、周纯禄、鲁美艳）

《哈尼语语法标注文本》（白碧波、许鲜明、邵丹）

《义都语语法标注文本》（李大勤、郭晓、宗晓哲）

《白语语法标注文本》（王锋）

《藏语甘孜话语法标注文本》（燕海雄、江荻）

《嘉戎语卓克基话语法标注文本》（林幼菁）

《壮语语法标注文本》（蓝利国）

《纳木兹语语法标注文本》（尹蔚彬）

《水语语法标注文本》（韦学纯）

《维吾尔语语法标注文本》（王海波、阿力木江·托乎提）

《达让语语法标注文本》（刘宾、孟佳仪、李大勤）

第一批著作的作者既是田野调查的实践者，又是调查描写的高手，他们把第一手材料用科学方法整合起来，费心尽力地加以标注，使得本套丛书展示出学术研究的深度和绚烂夺目的多样性族群文化色彩。对于年轻一代学者，包括在读博士生来说，尽管项

目仅要求基础标注和简短的语言导论,而语法单位的关联性和语法系统的体系性难度远超一般专题研究,给他们带来不小的挑战。他们记住了项目的目标和宗旨,即服务于学界,推动中国民族语言研究走向新的高度,开辟新的生长点和新的路径。我相信,这批著作的标注资源使得其他学科有了发力点,有了依托性,其价值之高怎么评价都不为过。在这个意义上,我也真诚呼吁中国最大的语言研究群体,广大的汉语研究学者,充分利用这个平台,巧用如此丰富的资源,透过你们的宏观视野和软实力,创造出更为恢宏的语言理论,甚或中国学者原创的学术体系。

当我初步编完这批著作,我由衷地相信,课题设计初衷所包含的另一个目的也已基本达成,这就是培养一批年轻学者。这个项目深化了他们的语言调查和专业分析技能,同时也推动他们创造出更多的优秀成果。

21 世纪初,中国学术界呈现出各学科的发展大势,总结 20 世纪的学术成就并预测新世纪的方向,中国民族语言学界也思考了民族语言研究的未来领域。我记得部分学者提出了民族语言的“本色语法”或“精深语法”研究,还有学者认为在全面的语言普查和初步描写之后应开展专题性深度研究,此外,语言材料的电子化典藏和文本资源的加工也是议题之一。现在,中国民族语言“语言志”(本色语法)项目已经启动,各语言专题研究已有不少成果,本丛书也初步实现了中国民族语言文本资源的熟语料汇集。不积跬步,无以至千里,不积小流,无以成江海,中国民族语言深度资源建设已上路。

江 荻

北京都会华庭寓所

2015 年 12 月 30 日

缩写符号

本文的语法标注符号根据《中国民族语言语法标注文本丛书课题语法标注集》选用，并据白语实际增加了个别标注符号。

1sg	First person singular	第一人称单数
1pl	First person plural	第一人称复数
1pl-incl	First person plural-inclusive	第一人称复数-包括式
1pl-excl	First person plural-exclusive	第一人称复数-排除式
1sg-POSS	First person singular-possessive	第一人称单数-领格
2sg	Second person singular	第二人称单数
2pl	Second person plural	第二人称复数
2sg-HON	Second person singular-honorific	第二人称单数-敬称
2sg-POSS	Second person singular-possessive	第二人称单数-领格
3sg	Third person singular	第三人称单数
3pl	Third person plural	第三人称复数
3sg-POSS	Third person singular-possessive	第三人称单数-领格
FEM/F	Feminine	阴性/雌性
MASC/M	Masculine	阳性/雄性
NOM	Nominalizer	名词化标记
ACC	Accusative	宾格
LOC	Locative	位格
POSS	Possessive	领有格
INCHO	Inchoative aspect	起始体
PROG	Progressive	进行体
PFV	Perfactive aspect	完成体
PEF	Perfect aspect	已行体
RST	Resultative	结果体
REP	Repetitive	反复体
TENT	Tentative	尝试貌
CAS	Casual	随意貌
PASS	Passive voice /-beed	被动态
REC	Reciprocal	相互态

AUX	Structural auxiliary word	结构助词
COMP	Complement particle	补语助词
ANP	Analogical particle	比拟助词
COOC	Coordinative conjunction	并列连词
ADVC	Adversative conjunction	转折连词
SEQC	Sequence conjunction	顺承连词
CAUC	Causal Conjunction	因果连词
NEG	Negative	实义否定
H-NEG	Have-NEG	存在否定
B-NEG	Be-NEG	关联否定
OUTW	Outward	外向方
INW	Inward	内向方
NSF	Non-selfhood	外向（朝向非我）
SEF	Selfhood (egocentricity)	内向（朝向自我）
IMPM	Imperative mood word	祈使语气词
IND	Indicative mood word	陈述语气词
INJ	Interjection mood word	感叹语气词
INT	Interrogative mood word	疑问语气词
JUD	Judging mood word	测度语气词
NEGM	Negative mood word	否定语气词
RHY	Rhythm mood word	节奏语气词
STIM	Sentence-initial mood word	句首语气词
TOP	Topic marker	话题标记

目 录

前 言 /1

缩写符号 /1

1 语法导论 /1

1.1 白语概况 /1

1.2 语音 /2

1.2.1 声母 /2

1.2.2 韵母 /4

1.2.3 声调 /7

1.3 词汇 /8

1.3.1 白语词汇的构成 /8

1.3.2 白语方言的基本词和方言词 /13

1.4 语法 /14

1.4.1 白语方言的词类及其语法特点 /15

1.4.2 白语的语序 /39

1.4.3 句类和句型 /41

1.5 文字使用情况 /49

1.5.1 老白文的流传和使用情况 /49

1.5.2 拼音白文的创制和使用情况 /51

2 标注文本 /53

2.1 大黑龙的故事 /53

2.2 洱岩洞 /58

2.3 弓鱼星的故事 /63

2.4 观音伏罗刹 /72

2.5 海神与海神姑娘 /85

2.6 华首门与洱海 /89

2.7 李珠告御状 /94

2.8 鲁川发文官武将 /97

2.9 青龙王和黄龙王 /107

2.10 三石三斗芝麻那么多的官 /117

2.11 石飞坪街 /125

2.12 杜朝选 /128

2.13 孤儿 /140

2.14 蝴蝶泉 /155

2.15 望夫云 /169

2.16 孝子 /189

2.17 阿黑和七姐 /198

2.18 阿金和三龙女 /209

2.19 大甌算和小甌算 /222

2.20 董法师 /230

2.21 段赤诚 /243

3 对照词汇 /259

后 记 /275