

基于不同文体的藏语名词统计研究

——以新闻与散文为例

赛毛吉

西北民族大学中国民族信息技术研究院, 甘肃 兰州

收稿日期: 2026年5月26日; 录用日期: 2026年7月17日; 发布日期: 2026年7月29日

摘 要

本文以藏语新闻和散文两种文体为研究对象, 采用语料库语言学与计量语言学方法, 系统考察了名词在两类文体中的形式特征、分布特征及语义差异。研究发现, 新闻文体中政治经济类复合名词占比显著, 呈现高信息密度特征; 而散文中自然意象与情感类单音节名词更为突出, 体现文学表达的具象化倾向。通过对比分析音节结构、词频分布、语义场分类等维度, 揭示了藏语文体差异对名词选择的深层制约机制, 为藏语信息处理、教学及跨文体翻译提供了数据支持与理论参考。

关键词

名词, 统计, 新闻文体, 散文文体

A Statistical Study on Tibetan Nouns Based on Different Textual Styles

—A Case Study of News and Prose

Maoji Sai

China Academy of Information Technology for Nationalities, Northwest Minzu University, Lanzhou Gansu

Received: May 26, 2026; accepted: July 17, 2026; published: July 29, 2026

Abstract

This paper takes Tibetan news and prose as the research objects and adopts the methods of corpus linguistics and quantitative linguistics to systematically investigate the formal features, distribution characteristics and semantic differences of nouns in the two types of texts. The research finds that in news texts, compound nouns related to politics and economy account for a significant proportion,

文章引用: 赛毛吉. 基于不同文体的藏语名词统计研究[J]. 现代语言学, 2026, 14(7): 1099-1104.

DOI: 10.12677/ml.2026.147743

presenting a high information density feature; while in prose, monosyllabic nouns related to natural images and emotions are more prominent, reflecting the tendency of concrete expression in literary expression. Through comparative analysis of syllable structure, word frequency distribution, semantic field classification and other dimensions, the paper reveals the deep-seated constraints of Tibetan text type differences on noun selection, providing data support and theoretical reference for Tibetan information processing, teaching and cross-genre translation.

Keywords

Nouns, Statistics, News Text Type, Prose Text Type

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

1.1. 研究背景

藏语作为汉藏语系的重要分支,拥有丰富多样的名词系统和文体传统。随着藏语信息处理技术的快速发展,对藏语词汇尤其是名词的量化研究需求日益增长。有学者对藏语名词进行了分类统计,分析了辞藻词、人名、地名和机构名称等专有名词的特征[1]。然而,现有研究多集中于单一文体的词汇分析(如新闻分词或文学词汇收集),缺乏对不同文体间名词使用差异的系统考察。在更广泛的文体研究领域,汉语及其他语言的文体差异研究已较为成熟,涵盖了从传统修辞学到计算语言学的多维探索。朱晓睿和骆健飞(2024)以词汇选择、句子长度等8项语言特征为正式度指标,尝试测量了12种不同文体的正式度与庄严度分数,探讨了汉语书面语正式度的测量方案[2]。谢永慧与杨尔弘(2024)则从量化词汇特征的角度出发,分别构建了新闻语体与文艺语体的等规模语料库,通过提取52个词汇层潜在特征并进行因子分析,揭示了两种语体在词汇层面的多维差异[3]。黄万武与李之祥(2025)基于布朗语料库,对新闻文体与小说文体的词汇频次进行了对比分析,发现两类文体在实词分布、词长及词汇密度等方面存在显著差异[4]。相比之下,藏语的类似研究仍处于起步阶段。文体作为语言功能的外在表现,深刻影响着词汇选择与组合方式。新闻与散文分别代表了实用型与文学型文体的两极:前者以信息传递为核心,后者以审美表达为旨归。二者在名词使用上必然存在显著差异,然而这种差异尚未得到充分的实证研究。

1.2. 研究目的与意义

本研究选取新闻与散文两类文体进行对比分析,具有多重意义。从理论层面看,填补了藏语文体学定量研究的空白,为建立藏语“文体词汇”对应模型提供基础数据;从应用角度看,研究成果可直接服务于藏语教学中的分文体词汇教学、机器翻译的领域适应优化以及跨文体转换的自动处理。尤其值得关注的是,藏语名词具有独特的形态特征(如丰富的词缀系统)和语义结构,在不同文体中可能呈现规律性变异。黄行等(2005)[5]曾从计算语言学角度对藏语名词组块的形式标记进行分类研究,但尚未涉及文体差异对名词使用的影响。

国内外相关研究现状显示,藏语词汇统计已取得一定进展。西北民族大学团队基于大规模语料库对藏语名词进行了分类统计,建立了专有名词识别词表;西藏大学在藏文文本分类研究中验证了名词特征对文体区分的重要性;另有学者对藏语形容词结构、词缀使用[4]等进行了专项研究。然而,这些研究或

限于单一文体内部分析，或未聚焦名词这一核心词类，更缺乏系统的跨文体对比。本研究将在现有成果基础上，采用更精细的计量标与对比方法，深入揭示藏语名词的文体分化规律。

2. 语料来源与研究方法

2.1. 语料来源与预处理

该散文语料共 9 万字，收录了 8 位当代藏语作家的代表作，其中，扎西达娃、次仁罗布、次仁顿珠、端智嘉是具有代表性的四位作家。扎西达娃开创了西藏魔幻现实主义风格，善于将神话传说与现实生活有机融合；次仁罗布注重从传统文学中汲取养分，以苦难与救赎的现实主义叙事，揭示现代文明与藏族传统文化之间的博弈；次仁顿珠善于书写藏区现代化进程中，社会与文化变迁背景下藏族人的生存境遇、情感世界及精神困境，其叙事深沉内敛，在悲剧氛围中展开人物生存环境、宗教信仰与传统习俗的多重冲突，以敏锐的探索意识冲击精神束缚；端智嘉被誉为藏族文坛的“奇才”，其创作涵盖诗歌、小说、散文等多种体裁，作品新颖独特且富有现实批判精神，有“藏族的鲁迅”之称，其诗作《青春的瀑布》开创了藏语自由诗的先河，对藏族当代文学具有开创性意义。该部分语料总体体现了文学藏语的审美特征。

语料预处理采用半自动化流程：首先使用西藏大学开发的藏文分词工具进行初始切分与词性标注，该工具基于 LSTM 和 CRF 模型；然后进行人工校验，重点处理分词歧义(如复合名词边界判定)和词性模糊(如动名兼类词)情况；最后建立带有文体标记的 XML 格式语料库。预处理阶段特别关注了名词的识别标准，除传统语法定义的名词外，包含名词性词缀(如“ལྔ”“ལྔ”等)的派生词也被纳入统计范围。

2.2. 研究方法

本研究采用定量分析和定性分析相结合的方法，从形式、分布和功能三个维度考察名词的文体差异。

形式特征方面，考察音节数量(单音节、双音节、多音节，其中多音节词按《藏文分词规范》细分为三至五音节及以上)、结构类型(单纯词、派生词、复合词)以及词缀使用频率。分布特征方面，通过相对频率(名词占文本总词数的比例)和重复率(同一名词在文本中的出现频次)两个指标进行量化。语义特征方面，基于语义场理论建立藏语名词分类体系，将名词划分为 12 个语义类别(如政治法律、自然景物、情感态度等)，统计各类别在新闻与散文中的分布差异，分类标准参考了藏语传统语法理论和现代语义学方法。最后，通过词频统计得到名词词汇在两种文体中的使用频次，并结合语言学相关理论分析名词的音节长度、构词特点和语义特征。

3. 名词的形式特征对比

3.1. 音节结构差异

通过对语料库的系统分析，发现新闻与散文在名词的音节分布上呈现显著差异(见表 1)。新闻文体中，双音节名词占比最高(58.7%)，其次为三音节及以上名词(31.2%)，单音节名词仅占 10.1%；而散文中单音节名词比例明显上升(24.5%)，双音节名词仍为主体(52.3%)，多音节名词下降至 23.2%。这种差异反映了两种文体的不同交际需求：新闻为追求信息表达的精确性与专业性，倾向于使用更多复合名词和术语；散文则应注重语言的形象性与节奏感，保留更多传统单音节词。

具体来看，新闻中高频多音节名词多属政治经济领域，如“དཔལ་ལྷན་ཁག་པོ་” dmangs gtso (民主)、“དཔལ་ལྷན་ཁག་པོ་ལྷན་ཁག་པོ་” dpal vbyor gong vphel (经济发展)等四至五音节复合词；散文中则常见“ཆུ་” chu (水)、“རི་” ri (山)、“ལྗང་” vod (光)等单音节基础词汇，这类词在文学传统中具有丰富的文化意象和情感联想。值得注意的是，两类文体在三音节名词的构成方式上也有区别：新闻多采用“名词 + 词缀”的派生形式

(如“slob gwa ba 学生”), 而散文更常用重叠式(如“skya rengs rengs 灰白”)构词法, 后者能增强语言的生动性。

Table 1. Comparison of syllable structure of nouns in news and prose (unit: %)
表 1. 新闻与散文名词音节结构对比(单位: %)

文体类型	单音节名词	双音节名词	三音节名词	四音节及以上名词
新闻(n = 12,358)	10.1	58.7	18.6	12.6
散文(n = 8742)	24.5	52.3	15.2	8.0

3.2. 构词方式差异

在构词类型方面, 新闻中复合名词占比高达 64.3%, 显著高于散文的 41.5% ($p < 0.01$)。如表 2 所示, 这些复合词在新闻中多采用修饰结构(如“vdzugs skrun 建设, 由“建立”建立与“制作”制作复合而成)和缩略形式(如“spyi tshogs 社会, 由“共同”共同与“集体”集体缩略合并); 散文中则多见派生词(如“rin po che 珍宝, 由“价”价与词缀“大”及“大”大派生而成)和形象复合词(如“gangs dkar 白雪, 由“雪”雪与“白”白复合而成), 后者体现了文学语言的具象化表达。

Table 2. Comparison of word formation types of nouns in news and prose
表 2. 新闻与散文名词构词类型对比

语料类型	词汇	构词类型	语素拆解	频次
新闻文体	建设	复合词	建立 + 制作	129
	社会	缩略词	共同 + 集体	309
散文文体	珍宝	派生词	价 + 大	125
	白雪	复合词	雪 + 白	56

词缀使用也呈现文体分化。新闻中名词性词缀(如表示职业的“”)出现频率为每千词 18.7 次, 散文中仅为 9.3 次。这与新闻需大量表达社会角色有关, 如“las byed pa (公务员)、“gsar vgod pa (记者)等。相比之下, 散文中形容词派生名词(如加“表示性质: “dkar po 白色)和小称词缀(如“vu 表亲昵: “byi vu 小鸟)使用更多, 反映文学语言的情感色彩。

4. 名词的语义特征对比

4.1. 高频名词分析

通过提取的两类文体中的名词并进行统计, 结果显示明显的领域分化。如表 3 所示, 新闻与散文在前 10 个高频名词的词汇选择上已呈现出迥异的语义取向。进一步分析前 100 个高频名词的语义分布, 如图 1 和图 2 所示, 新闻中出现频次最高的 100 个名词里, 76%属于政治经济(如“khrims lugs 法律)、社会管理(如“gzhung vberl 政府关系)和科技术语(如“tshan rig lag rtsal 科学技术); 散文中则有 68%的高频名词属于自然景物(如“gangs 雪山)、情感体验(如“dgav-spro 喜悦)和文化符号(如“mchod rten 佛塔)。

专有名词的使用也体现文体差异: 新闻中地名(如“lha sa 拉萨)、机构名(如“mi dmangs khrims khang 人民法院)占比达 22.4%, 且多采用全称形式(如“bod rang skyong ljongs 西藏自治区); 散文中人名(如作家创造的“bstan vdzin 丹增)、文学地名(如“sman ri 药山)

Table 3. Top 10 most frequent nouns in news and prose corpora**表 3.** 新闻文体和散文文体语料中出现的前 10 个频次的名词词汇表

新闻文体高频名词	拉丁转写	频次	散文文体高频名词	拉丁转写	频次
མི་དམངས་ (人民)	dmangs gtso	462	གངས་ (雪山)	gangs	379
དོན་ (管理)	do dam	447	མཚོ་ (湖)	mtsho	371
དཔལ་འབྱོར་ (经济)	dpal vbyor	412	ལོ་ (光)	vod	357
ཚོན་ལས་ (产业)	thon las	372	སེམས་ (心)	sems	355
ཁྲིམས་ལུགས་ (法律)	khirms lugs	323	དགའ་སྡེ་ (喜悦)	dgav spro	347
ཚན་རིག་ལག་རྩལ་ (科技)	tshan rig lag rtсал	316	མེ་དྲག་ (花)	me tog	345
ཕྱི་ཚོགས་ (社会)	spyi tshogs	310	ཆུ་ (水)	chu	342
གུང་གི་ (中国)	krung go	309	ཆུ་ (灰色)	skya	340
སློབ་གསུང་ (教育)	slob gso	276	ཟླ་བ་ (月亮)	zla ba	337
དམངས་ཚོགས་ (民众)	dmangs tshogs	265	རྩ་ (鼓)	rnga	326

**Figure 1.** Word cloud of the top 100 most frequent nouns in news**图 1.** 新闻名词出现的前 100 个频次词云图**Figure 2.** Word cloud of the top 100 most frequent nouns in prose**图 2.** 散文名词出现的前 100 个频次词云图

仅占 8.7%，且常使用亲昵变体(如“*ལྷ་ས་ཕུག་*”拉萨小城)。

4.2. 语义场分布差异

基于建立的 12 类语义场系统，统计发现新闻中“政治法律”(32.1%)、“社会经济”(28.7%)和“科学技术”(11.2%)三类名词占比超过 70%；散文中“自然景物”(38.5%)、“情感态度”(22.3%)和“文化宗教”(19.8%)占主导地位。这种分布与文体功能高度吻合：新闻需客观报道社会运作，散文则主观表达个体体验。

在某些语义场内部也存在文体分化。如同样表达“人”的概念，新闻多用角色名词(如“*གསར་འགྱུར་པ་*”记者)、职务称谓(如“*ཡུལ་ཤེས་*”主任)；散文则倾向使用关系称谓(如“*མ་མ་*”a ma 母亲)、特征描述(如“*སྐྱོན་*”忧郁者)。在自然类名词中，新闻侧重资源性词汇(如“*ནགས་ཚལ་*”nags tshal 森林)，散文偏爱审美性词汇(“*ནམ་མཁའ་*”nam mkhav 天空)。

5. 结论与应用展望

本研究通过系统的定量分析，揭示了藏语名词在新闻与散文两种文体中的差异化分布规律。其一，形式特征上，新闻名词呈现明显的“多音节化”和“复合化”倾向(双音节及以上占 89.9%，复合词占 64.3%)，散文则保留了更多传统单音节词(24.5%)和单纯词(58.5%)，反映了两种文体对信息密度与审美表达的不同需求。其二，语义分布方面，新闻名词高度集中于政治经济、社会管理等“刚性”语义场(占比超过 60%)，散文名词则聚集于自然景物、情感体验等“柔性”领域(占比超过 80%)，印证了 Biber 的文体维度理论[6]：新闻属于“信息性”文体，散文属于“互动性”文体。其三，文化内涵层面，新闻名词体现了现代藏语的标准化趋势(如大量使用政治术语和科技术语)，而散文名词则承载了更多传统元素(如宗教文化词汇和自然意象词)，反映了藏语在现代化进程中“标准语”与“文学语”之间的张力。在应用层面，本研究呈现的词汇分布特征对藏语阅读教学中文体意识的培养有一定参考价值，但研究未涉及名词的搭配模式与具体语用功能，尚不足以支撑简化的词汇教学方案。此外，研究所观察到的现代藏语名词发展趋势(如术语复合化、词缀规范化)可为藏语新词术语制定提供基础数据，散文中保留的传统词汇也可作为藏语文化遗产保护的参考资料。

参考文献

- [1] 祁坤钰. 基于语料库的藏语名词分类与统计研究[J]. 西北民族大学学报(自然科学版), 2012, 33(3): 44-49.
- [2] 朱晓睿, 骆健飞. 汉语正式度测量与文体差异研究[J]. 国际中文教学资源研究, 2024(1): 143-160.
- [3] 谢永慧, 杨尔弘. 新闻语体与文艺语体词汇特征差异的多维定量分析[J]. 语言教学与研究, 2024(3): 68-78.
- [4] 黄万武, 李之祥. 基于布朗语料库的新闻文体与小说文体的词汇频次对比[J]. 中文科技期刊数据库(文摘版)社会科学, 2025(1): 7-9.
- [5] 黄行, 孙宏开, 江荻, 等. 现代藏语名词组块的类型及形式标记特征[C]//南京师范大学, 清华大学智能技术与系统国家重点实验室. 全国第八届计算语言学联合学术会议(JSCL-2005)论文集. 北京: 清华大学出版社, 2005: 627-629.
- [6] Biber, D. (1988) Variation across Speech and Writing. Cambridge University Press.
<https://doi.org/10.1017/CBO9780511621024>