文章编号：2095-2163(2019)01-0131-04 中图分类号：TP391.1 文献标志码：A

方块苗文词性标注集的设计

周 潭，莫礼平，曾 虎，雷 智，李文宇，吴 莹
（吉首大学 信息科学与工程学院，湖南 吉首 416000）

摘 要： 词性标注集是计算机处理自然语言时进行词类表示的工具。任何自然语言的词性标注都必须以词性标注集为基础。本文根据方块苗文信息化的实际需要，结合方块苗文的造字原理及词语的使用特点，在介绍对词性标注及标注集相关概念的基础上，参考汉语词性标注规范设计方法，基本确定了方块苗文的词性和种类，设计了方块苗文的词性标注符号和基于语法范畴的分类标记体系；初步制订了用于方块苗文信息处理领域的词性标注集，在某种意义上为方块苗文词性标注建立了参考标准。

关键词： 自然语言处理；方块苗文；词性标注；词性标注集

Design of the part-of-speech tag set for the square Hmong characters

ZHOU Tan, MO Liping, ZENG Hu, LEI Zhi, LI Wenyu, WU Ying
(College of Information Science & Engineering, JiShou University, Jishou Hunan 416000, China)

【Abstract】 The part-of-speech (POS) tag set is a tool for word class representation when the computer processes natural language. The POS tagging of any natural language must be based on a set of POS tags. According to the actual needs of the informatization of the square Hmong characters, combined with the word-making principle and the use characteristics of the words, the POS and type of the square Hmong characters are basically determined by referring to the design method of Chinese POS tagging specification. And then, the POS tagging symbol and the classification tagging system based on the grammatical category are designed. A more complete POS tagging set for the square Hmong characters information processing field is preliminarily developed. In a certain sense, the reference standard of the POS tagging for the square Hmong characters is established.

【Key words】 natural language processing; square Hmong character; part-of-speech tagging; part-of-speech tag set

0 引 言

词性标注是自然语言信息处理的基本内容，也是文本索引、文本分类、语言合成、语料加工、机器翻译、信息检索等应用领域不可缺少的一个环节。从概念上来讲，词性标注就是根据上下文语法关系，判定句子中每个词语的语法范畴，以确定其词性并加以标注的过程[1]。词性标注对于消除词语歧义、减少查询模糊性、降低索引量、提高检索效果及效率有着重要作用。Wilks Y[2]在利用英文词性标注帮助实现语义消歧的研究中指出“词性标注是实现语义消歧的第一步”；利用词性标注减少查询模糊性的研究也表明，在搜索引擎的信息查询系统中加入词性标注功能，有助于系统更准确理解用户的查询意图，能够返回更符合用户要求的结果[3]。词性标注集是用来表示词类的工具，是词性标注系统必不可少的组成部分。任何语言的词性标注都必须以词性标注集为基础。词性标注集的建立及词性标注技术的发展与语料库建设紧密相联。20 世纪 60 年代初，以 N.Francis 为首的一批学者在美国的布朗大学建成了当今最早的机读语料库—Brown 语料库。在 Brown 语料库的建设过程中，人们开始研究英文词性标注问题。20 世纪 80-90 年代，进入第二代、第三代电子语料库时代，百万、千万、上亿词级的深度标注语料库出现。此期间，英语词性标注集的建立及词性标注技术得到了长足的发展[4]。历经近 50 多年的发展，英语词性标注集的建立及词性标注技术已趋于成熟，面向德文[5]、法文[6]、意大利文[7]、阿拉伯文[8-9]、印度文[10]的各种类型文本词性标注集的建立及词性标注技术也发展迅速。近年来，面向稀缺语言[11]的词性标注研究工作逐渐见诸报道。

本文根据方块苗文这一稀缺语言文字的造字原

基金项目：国家自然科学基金(61462029)；吉首大学本科生科研项目(JDX17027,2018JDX09)；大学生研究性学习和创新性实验计划项目(湘教通〔2018〕255 号文件,599；吉首大学教通 2018[15]号文件,JDCX2018012)。

作者简介：周 潭(1998-)，男，本科生，主要研究方向：自然语言处理；莫礼平(1972-)，女，硕士，高级实验师，主要研究方向：自然语言处理、Petri 网理论及应用。

通讯作者：莫礼平 Email：zmx89@163.com

收稿日期：2018-09-18

理及方块苗文词语的使用特点,结合对方块苗文原始语料中词语统计的结果,讨论方块苗文词语的种类、词性的确定和划分方法,并尝试参考汉语词性标注规范及标注集的设计方法,制订方块苗文的词性标注符号和词性标记集。

1 方块苗文的造字原理及使用特点

湘西方块苗文是清朝末年以来武陵山地区湘西苗族人民根据"取个人认为最易认、易记的汉字作为代表符号"的原则自创的一种文字[12-13],是中国民间苗族文化的主要载体之一。湘西方块苗文分为3种,按照其产生和使用地区,分别称之为板塘苗文、老寨苗文、古丈苗文。这3种苗文都是清末以来苗族文人为记录、整理、创作苗歌而创造的汉字式的苗文。当地苗族群众称之为"土字"、"乡字"。其中,板塘苗文至今仍在附近数百里苗乡流传使用,对苗族文化的发展和传播起到了良好的推进作用。

湘西方块苗文是武陵山地区湘西苗族人民因民族文化生活的需要而产生的一种借源文字。3种苗文均脱胎于汉字这一母体,记录同一种语言。因此,3种方块苗文在结构和造字法方面都不谋而合,均采用方块结构,且基本上都是一字一音节地标记一个语素或词;并借鉴了汉字的造字方法,创造性地运用形声、会意、假借、双声、象形等手段,借用包括义符、声符、形符(特指象形偏旁)在内的汉字或汉字构件造字,酷似汉字而实非汉字。方法苗文中,源自形声构字的湘西方块苗文约占总字数的四分之三左右[12-13]。不同结构的方块苗文及其汉义如图1所示。

结构类型	字例	汉义
左右型	猡 貔	猪 猪
上下型	蜡 虫	认识 蛇
侧围型	𠀧 凸	一个 我们
内外型	闯 闯	出去 门

图1 不同结构的方块苗文及其汉义

Fig. 1 Examples of square Hmong characters in different structure and the corresponding Chinese meanings

实际应用中,方块苗文通常是在苗族人民创作的歌本、剧本中与汉字混合出现。方块苗文的词语以单字词为主,有少量双字词,鲜有3字及以上的词语出现。

2 词性标注相关概念及经典词性标注集

2.1 相关概念

词性(part-of-speech)又称为词类,是以语法特征(包括句法功能和形态变化)为主要依据、兼顾词汇意义的基础上,对词进行划分的结果。从组合和聚合关系来说,一个词类就是指众多具有相同句法功能,并可以在同样的组合位置中出现的词聚合在一起所形成的语言范畴。词类划分具有层次性。例如,汉语中的词可以分成实词和虚词,实词中又包括体词、谓词等,体词中又可以分出名词和代词等。

词性标注(part-of-speech tagging),又称为词类标注,是指为分词结果中的每个单词标注一个正确的词性的过程,也即确定每个词是名词、动词、形容词或者其它词性的过程。

词性标注集系指对词的类别进行划分的集合,是用来表示词类的工具,是词性标注系统必不可少的组成部分。某一种语言的一套词性标注集应该将词性具体划分为多少个类别,没有统一的规定。因此,同一种语言可能建立多种词性标注集。在语言的实际应用中,使用什么样的词性标注集取决于应用的目的及应用所需要的信息量。在计算语言学不同的领域中,人们对词性依赖的程度不同,处理的精度不同,所以对词性分类的粒度的定义也不同。

2.2 经典词性标注集及词性编码

词性标注有小标注集和大标注集。例如,小标注集将所有代词都归为一类,而大标注集则将代词进一步细分为指示代词、人称代词和疑问代词3类。采用小标注集比较容易实现,但是太小的标注集可能会导致类型区分度不够。划分越细致的词性标注集越有利于信息区分,但词性标注的难度也就越大。

每种语言都有自己的词性标注集。最早出现的词性标注集是英文词性标注集,著名的Brown语料库所使用的标注集(Brown标注集)是历史上最有影响的英语标注集,包含87个标记。多数的英语标注集都是从该标注集发展而来。宾州树库(Penn Treebank)使用的标注集是经过对Brown标注集的简化而得到,包括45个标记,现在已经成为计算语言学领域使用最为广泛的英语标注集。同英语词性标注集一样,汉语词性标注集目前也没有统一的标准。因此,根据对词性分类体系的不同理解和不同应用领域的要求,出现了多种汉语词性标注集。中科院计算所刘群等制订的"ICTPOS3.0汉语词性标记集"是当前最有影响的词性标注集之一。

为了方便指明词的词性,词性标注集需要给每个词性编码。例如,见表1的《PFR人民日报标注语料库》的词性编码表就是根据词性的英文单词的首字母进行词性编码。把"名词"编码成"n"、"形容

词”编码成“a”、“动词”编码成“v”等。

表 1 《PFR 人民日报标注语料库》的词性编码表截表

Tab. 1 Screenshot of the POS code list of "PFR People´s Daily Labeling Corpus"

词性编码	词性名称	注解
Ag	形语素	形容词性语素
a	形容词	取英语形容词 adjective 的第 1 个字母
ad	副形词	直接作状语的形容词
an	名形词	具有名词功能的形容词
b	区别词	取汉子“别”的声母
c	练词	取英语连词 conjunction 的第一个字母
d	副词	取 adverb 的第二个字母
dg	副语素	副词性语素
e	叹词	取英语叹词 exclamation 的第一个字母
f	方位词	取汉字“方”
g	语素	绝大多数语素都能作为合成词的“词根”
h	前接部分	取英语 head 的一个字母
i	成语	取英语成语 idiom 的第一个字母
j	简称略语	取汉字“简”的声母
…		
un	未知词	不可识别词及用户自定义词组

3 方块苗文词性分类及词性标注集的设计方案

3.1 方块苗文的词类、词性

仿汉字结构的方块苗文基本上是一字标记一个语素或词。因此，实际应用文档中出现的方块苗文词语主要是单字词和少量双字词，极少有 3 字及以上的词语。这些词语的词性及使用方法类似于汉语词语。而且，相对于汉语，方块苗文的词语的数量较少，词性也相对较少，语法较为简单，出现兼词的词语数量也很少。

在对已整理出的苗文初级生语料进行统计分析的基础上，根据方块苗文词语的汉义，将方块苗文词语分为实词和虚词 2 个大类。实词分为 7 个小类：名词、动词、形容词、数词、量词、代词、副词。虚词分为介词、助词 2 个小类。此外，针对个别不可识别的词语以及当前汉义不明确的词语，增加一个类别，即未知词。

3.2 方块苗文词性标注集的设计方案

由于方块苗文词语是根据其汉义确定所属词性分类，故可以直接借鉴汉语词性标注集的词性编码方法进行方块苗文词性标注集中词性标注的符号、形式和风格的设计。依据北京大学计算语言学研究俞士汶主编的《现代汉语语料库加工—词语切分与词性标注规范与手册》，研究制订了方块苗文词性标注集，该词性标注集的部分内容见表 2。

表 2 方块苗文词性标注集中的词性编码表截表

Tab. 2 Screenshot of the POS code list in the POS tag set for the square Hmong characters

词性编码	词性名称	注解
n	名词	取名词对应英语形容词 noun 的第一个字母，如“[illegible]、[illegible]、[illegible]”，汉义分别为“小山丘、田、石头”
a	形容词	取形容词对应英语 adjective 的第一个字母，如“[illegible]、[illegible]、[illegible]”，汉义分别为“暗的、长的、短的”
an	名形词	具有名词功能的形容词，如“[illegible]”，形容词性义为“老的、旧的”，名词词性汉义为“历史、故事”
f	方位词	取汉字“方”拼音的首字母，如“[illegible]、[illegible]、[illegible]、[illegible]”，汉义分别为“东方、南方、后方、前方”。
m	数词	取英语 numeral 的第三个字母，如“[illegible]、[illegible]、[illegible]”，汉义分别为“三、四、百”。
v	动词	取英语动词 verb 的第一个字母，如“[illegible]、[illegible]、[illegible]”，汉义分别为“喊、怕、结冰”。
p	介词	取英语介词 preposition 的第一个定母，如“[illegible]”，汉义为“别，不要”、“勾”，意为“拿、把、用”
d	副词	取 adverb 的第二个字母
td	时间副词	取 time 的第一个字母，adverb 的第二个字母，如“[illegible]”，汉义为“一阵子”
q	量词	取英语 quantity 的第一个字母
r	代词	取英语 pronoun 的第二个字母
t	时间词	取英语 time 的第一个字母
uv	未知词	不可识别词及用户自定义词组
sy	符号词	Symbol 的前 2 个字母（如“X”“~”）

4 方块苗文词性标注示例

依据上述的方块苗文词性标注集，笔者尝试使用词性标注算法对整理出的苗文手稿资料文档中的部分方块苗文进行词性标注。标注效果如图 2 所示。

5 结束语

方块苗文是武陵山地区民间苗族文化的主要载体之一，其信息化对于弘扬苗族文化，推进民族文化旅游产业发展和苗族文化非物质遗产数字化保护进程有着重要意义。然而，方块苗文信息处理研究起步较晚，仅有莫礼平等人[14-18] 2013 年以来，在方块苗文的字信息处理层面取得的少量研究成果见于报道。实现字信息处理层面基本技术之后，语信息层面的词性标注技术研究和词性标注语料库建设成为方块苗文信息化研究需要解决的问题。而设计词性标注规范，并据此制订完整的词性标记集，正是实现方块苗文词性标注技术和建设方块苗文语料库的重要环节。本文结合方块苗文的造字原理及词语的使用特点，参考汉语词性标注规范及标注集的设计方法，初步制订的方块苗文词性标记集，将成为方块苗文词性标注过程中进行词类表示的参考工具，为最终完整的信息处理用方块苗文词性标记集的建立打下良好基础。

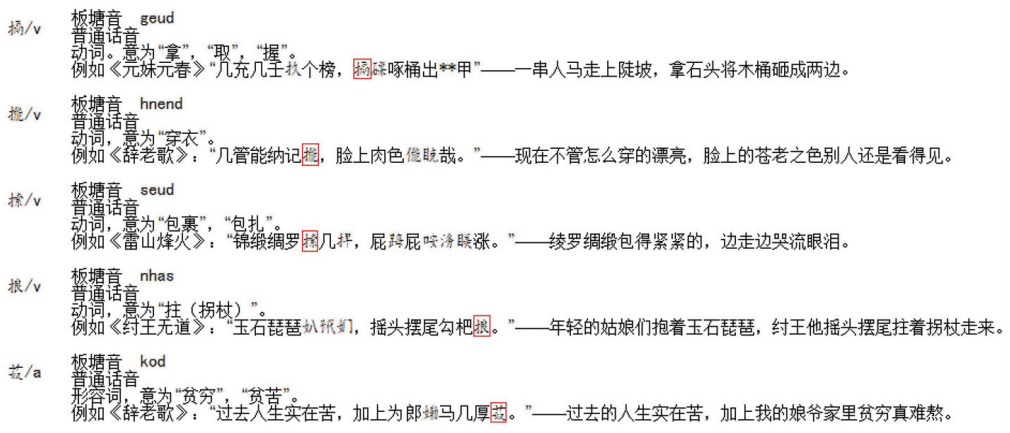

图 2 词性标注效果截图

Fig. 2 Screenshot of the POS tagging effect

参考文献

[1] MANNING C D, SCHUTZE H. Foundations of statistical natural language processing[M]. London: the MIT Press, 2000.

[2] WILKS Y, STEVENSON M. The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation[J]. Natural Language Engineering, 1998, 24(1): 1-9.

[3] ALLAN J, RAGHAVAN H . Using part-of-speech pattern to reduce query ambiguity[C]//Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. Tampere, Finland: ACM, 2002: 307-314.

[4] LEECH G. The state of the art in corpus linguistics[M]//AIJMER K, ALTENBERG B . English CorpusLinguistics: Studies in Honor of Jan Swartvik . London: Longman, 1991: 18-29.

[5] WESTPFAHL S, SCHMIDT T. FOLK-Gold — A gold standard for part-of-speech-tagging of spoken German[C]//Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia: dblp, 2016: 1493-1499.

[6] CHRISTODOULIDES G, AVANZI M, GOLDMAN J P. DisMo: A morphosyntactic, disfluency and multi-word unit annotator. An evaluation on a corpus of French spontaneous and read speech[J]. arXiv preprint arXiv: 1802.02926, 2018.

[7] CRISTINA B, TAMBURINI F, BOLIOLI A, et al. Overview of the EVALITA 2016 part of speech on twitter for Italian task[C]// CEUR Workshop Proceedings (CEUR-WS.ORG). SANTIAGO, CHILE: CEUR, 2016, 1749: 1-7.

[8] ABUMALLOH R A, AL-SARHAN H M, IBRAHIM O, et al. Arabic part-of-speech tagging[J]. J. Soft Comput. Decis. Support Syst, 2016, 3(2): 45-52.

[9] ZEROUAL I, LAKHOUAJA A, BELAHBIB R. Towards a standard part of speech tagset for the Arabic language[J]. Journal of King Saud University - Computer and Information Sciences, 2017, 29(2): 171-178.

[10] SARKAR K. Part-of-speech tagging for code-mixed Indian social media text at ICON 2015[J]. arXiv preprint arXiv: 1601.01195, 2016.

[11] KIM Y B, SNYDER B, SARIKAYA R. Part-of-speech taggers for low-resource languages using CCA features[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015: 1292-1302.

[12] 赵丽明，刘自齐. 湘西方块苗文[J]. 民族语文，1990(1)：44-49.

[13] 杨再彪，罗红源. 湘西苗族民间苗文造字体系[J]. 吉首大学学报(社会科学版)，2008，29(6)：130-134.

[14] 莫礼平，周胜，尹楠佳. 基于构件汉语拼音的湘西方块苗文输入法[J]. 吉首大学学报(自然科学版)，2014，35(2)：30-34.

[15] 莫礼平，曾水玲，周恺卿. 音形结合的方块苗文输入编码方案研究[J]. 计算机科学与探索，2014，8(8)：1017-1024.

[16] 莫礼平，周恺卿. 一种湘西民间苗文字形的动态生成方法及其实现途径[J]. 北京大学学报(自然科学版)，2016，52(1)：141-147.

[17] 莫礼平，周恺卿，蒋效会. 基于 OpenType 技术的方块苗文字库研究[J]. 中文信息学报，2015，29(2)：150-156.

[18] 曾磊，莫礼平，刘笔余，等. Horspool 扩展算法在方块苗文模式匹配中的应用[J]. 吉首大学学报(自然科学版)，2018，39(4)：32-37.

(上接第 130 页)

参考文献

[1] 路阳，付艳明，张卯瑞. 线性连续周期系统的模型参考跟踪控制[J]. 控制与决策，2016，31(7)：1279-1284.

[2] 李维鹏，张国良，姚二亮，等. 基于空间位置不确定性约束的改进闭环检测算法[J]. 机器人，2016，38(3)：301-310，321.

[3] 刘聪，李颖晖，刘勇智，等. 采用高阶终端滑模观测器的执行器未知故障重构[J]. 西安交通大学学报，2015，49(9)：126-133.

[4] 曾爱林. 基于 Android 的心电实时监护系统设计与实现[J]. 计算机测量与控制，2013，21(11)：2997-3000.

[5] 黄德军，贾如春，李林原. 基于 Web Services 的 SOA 自动化控制架构的研究与实现[J]. 自动化与仪器仪表，2018(7)：63-67.

[6] 高泽仁，周丰，赵庆佳，等. 光电滑环性能动态检测技术研究[J]. 中国电子科学研究院学报，2018，13(2)：174-180.

[7] 章宝歌，田铭兴. 新型磁控电抗器特性谐波研究[J]. 电源学报，2018，16(3)：168-172.

[8] 戴立坤. 大型多媒体网络通信中的安全监测平台设计[J]. 现代电子技术，2016，39(24)：9-13.

[9] 陆兴华，甄汉健，段五星. 嵌入式多模控制系统的容错性控制器设计[J]. 机械与电子，2016，34(4)：62-65.

[10] 鞠晓东，郑振. 基于末端轨迹修正的导弹跟踪稳定性控制方法[J]. 智能计算机与应用，2018，8(2)：121-125.