侗台语族语言的编辑距离分类

赵志靖[1],江 荻[2]
ZHAO Zhijing[1],JIANG Di[2]

1.扬州大学,江苏 扬州 225009
2.中国社会科学院,北京 100081
1.Yangzhou University, Yangzhou, Jiangsu 225009, China
2.Chinese Academy of Social Sciences, Beijing 100081, China

ZHAO Zhijing, JIANG Di. Classification of levenshtein distance of Dong-Tai language family languages. Computer Engineering and Applications, 2018, 54(19):62-67.

Abstract: The levenshtein distance is a distance metric derived from the number of edit operations needed to transform one string into another. This metric has received recent attention in Western countries as a means of automatically classifying languages into genealogical subgroups, and has been proved to be effective in the measurement of the distances between languages or dialects. This paper applies the algorithm of the levenshtein distance to the computational classification of the Dong-Tai language family languages, and their genetic relationship is described. The calculation results show that the language classification of the levenshtein distance is consistent with that of the historical linguistics, and a new way is proposed for the computational method. The levenshtein distance can be applied to the research of the East Asian languages.
Key words: Dong-Tai language family; levenshtein distance; language classification

摘 要: 编辑距离是一种距离测量法,源于将一个字符串变换为另一个字符串所需要的编辑操作数,该方法能够自动将语言进行分类,最近这些年在西方很受关注,被证明测量语言或方言间距离是有效的。运用编辑距离算法对侗台语族语言做出计量分类以及亲缘关系程度的描述。结果表明编辑距离分类结果与历史语言学的分类结果是基本一致的,为计量法提供了新思路。编辑距离可以应用于东亚语言的研究中。
关键词: 侗台语族;编辑距离;语言分类
文献标志码: A **中图分类号:** H087;TP39 doi:10.3778/j.issn.1002-8331.1708-0392

1 侗台语族语言的传统分类

李方桂将侗台语族分为两大语群,即台语群(壮语次群,西南次群(泰、傣语等))和侗水语群(侗语次群,水语次群,莫语次群,佯黄语次群),临高话属于壮语次群,没有定黎语群。

罗常培将中国境内的侗台语族分为3个语支,即壮傣语支(壮语、布依语、侬语、沙语、傣语),侗水语支(侗语,水家话(毛南、莫家、佯璜的语言看做水家语的方言))和黎语支(黎语)。

1987年《中国语言地图集》将侗台语族14种语言分为3个语支,即壮傣语支(壮语、布依语、傣语、临高话),侗水语支(侗语、水语、仫佬语、毛南语、佯璜语、莫语、拉珈语),黎语支(黎语和村话),此外仡佬语是否作为语支未定。

梁敏、张均如建立了一个与黎、侗水、台平行的仡央语支,将侗台语族分为4个语支,即台语支(包括国内的壮语、布依语、傣语、临高话和国外的泰语、老挝语、掸语、石家话、土语、侬语、岱语、黑泰语、白泰语、坎梯语和已趋于消亡的阿含语等),侗水语支(包括侗语、仫佬语、水语、毛南语、莫语、锦语、佯璜语、拉珈语、标语等),黎语支(包括黎语、村语),仡央语支(包括仡佬语、拉基语、普标语、布央语、耶容语和越南北部的拉哈语等)。

基金项目: 教育部人文社会科学研究青年基金(No.15YJC740141);江苏高校哲学社会科学研究项目(No.2015SJB783)。
作者简介: 赵志靖(1978—),男,博士,讲师,研究领域为计算语言学,E-mail:zhaojsj@163.com;江荻(1954—),男,博士,教授,研究领域为汉藏语言学和计算语言学。
收稿日期: 2017-08-31 **修回日期:** 2017-10-26 **文章编号:** 1002-8331(2018)19-0062-06
CNKI **网络出版:** 2018-03-22, http://kns.cnki.net/kcms/detail/11.2127.TP.20180321.1706.016.html

本尼迪克特(P.K.Benedict)提出了卡岱语，分别将仡佬、黎语、临高、拉嘉在3个不同的层次上独立。

本文借助计算机手段，基于斯瓦迪士100核心词，运用编辑距离算法对侗台语族16种语言进行分类。

2　侗台语族语言的编辑距离计算

2.1　比较词表的选择

确定语言分类的时候，选择比较词项是一个很关键的问题。择词合理与否直接影响到比较的结果。由于词的性质不尽相同，同时比较词项又有数量上的要求，因此操作起来十分困难。这就涉及到了可供操作且符合比较目的的比较词表的选择问题。选择多少数目的关系词才较适合作语言关系的分类比较，这个问题很重要。

东亚语言历史研究中确定同源词一直是一个费解的难题，学者们花费了大量精力构建各类破解方法，企图甄别同源词与借词，以达到判断语言相互关系的目的。江荻[1]认为“各种研究方法都不同程度深化和逼近了研究目标，但是各种方法又都有局限性。……所以我们又回到了甄别同源词与借词的原点”。另外，还有学者利用构造适合东亚语言比较的小规模核心词集来简化问题[2]，如Matisoff建立的东南亚语言的200词词表；黄布凡提出的300词的藏缅语核心词词表；郑张尚芳建立的华澳语言比较300词的词表；江荻提出的200词的汉藏语核心词表。江荻认为[1]“这些核心词集基本都是经验性的，缺乏可信的选词理据，同时，这些词集很少得到应用，难以判断实际效用。”以上词表都是模仿斯瓦迪士核心词集，期望寻找适合汉藏语言的核心词集。这些研究基础不同，所采用的词汇标准大相径庭，得出的结论自然不同，主观性很强。

美国学者Swadesh为计算词汇反映的史前民族接触深度，提出了语言年代学概念及相关公式和方法，并创造了一个最具普遍性的200词表(后修改另设100词表)。他所提出的词表得到印欧语等多种语言历史年代分化数据的间接验证，具有实践应用经验。很多学者也都利用斯瓦迪士词表做相关研究，如日本学者王育德1962年用斯瓦迪士200词表计算汉语方言北京话、苏州话、广东话、梅县话和厦门话之间的关系[3]；徐通锵先生在1991年将斯瓦迪士100核心词表应用到语言年代学的计算中，计算出了汉语方言北京话、苏州话、长沙话、南昌话、广州话、梅县话、厦门话之间的同源百分比和分化年代[4]；梁敏利用斯瓦迪士200词表对仡佬、拉基、普标、布央等语言进行了研究，提出“为了避免选词时的主观倾向，以斯瓦迪士有关语言年代学统计中所采用的包括200多个基本语词的词表作基础，从中剔除那些在我们对比的语言中没有的或用词组表示的词项，最后选定了200个词项作对比的基数(在某些语言之间也可能不足200个)”[5]；王士元用斯瓦迪士百词表划分了侗台语的谱系树[6]；毛宗武、李云兵用斯瓦迪士修正100词表和基本200词表，将炯奈语与苗瑶各语言或方言互相比较[7]；Oswalt、Guy、Ringe、Kessler、Goh、Brown等利用斯瓦迪士100词表对语言进行分类[8]；德国马普所的ASJP项目采用斯瓦迪士100词，后来又采用斯瓦迪士100词中的40词对语言进行自动分类[9]；陈保亚[10-11]利用斯瓦迪士的第100核心词与第200核心词比例来观察语言或方言之间的关系，经过他的广泛应用，产生了词集分层次的高低阶概念，催生了关系词阶曲线判定法。关系词阶曲线判定法已取得令人满意的成果，已可初步判定相关语言的关系。认为百词表比语法、语音系统更稳定，不易借用；孙宏开用斯瓦迪士100词基本词表，以滚董话代表巴哼语，将苗瑶各语言或方言互相比较[12]；邓晓华、王士元[13]提到“斯瓦迪士的基本词汇表已成功适用于世界上的多种语言，例如‘罗赛塔计划’。”“国内大多数语言学者过分强调汉藏语言的特殊性，自立一套词表，忽略斯瓦迪士百词表的国际性、可比性和计量原则”他们利用略有调整的斯瓦迪士100核心词(主体仍是斯瓦迪士100核心词)对苗瑶语族、藏缅语族和壮侗语族做了计量分类；江荻[1]用基本层次范畴理论构建核心词范畴以及为核心词范畴择词，择词以斯瓦迪士核心词为来源，观察各词项进入范畴和满足基本层次范畴的隶属程度要求，增补删减，构建出修订的斯瓦迪士核心词集；江荻[14]采用词频统计的方法观察斯瓦迪士词表的分布特征，然后提出以词频方法构建核心词表。

斯瓦迪士词表是在印欧语调查研究的基础上，经过反复的实践而筛选出来的，具有普遍性，比较稳定。它的借用率很低，衰变率在不同的亲属语言中基本是相同的，用百词表中同源词比例的高低来确定同源语言亲属关系的远近比其他方法似乎更可靠。从核心对应语素的比例来划分谱系树更能够排除语言借用的干扰。两种同源语言百词表中同源词数量越多，它们的亲缘关系越近。尽管斯瓦迪士词表的适用性和可用性存在争议，但在世界范围语言历史研究中获得了广泛的应用，对世界各地语言具有一定的普适性，被各界学者广泛运用来比较语言/方言之间的亲属关系，至少是目前国际语言学界公认的做历史语言学比较的最佳优选词目，同时具有比较强的可操作性。由于目前国内外还没有人拿出更合理、更有说服力、实践性更强的核心词表，这本身也是一项非常困难的工作，因此本文计算的对象选用斯瓦迪士的100核心词。斯瓦迪士100核心词不是本文主观拟定的，因此具有反映研究目的的效度。正如徐丹所说[15]“在语言学者没有其他更好的方法之前，这一词表仍然被广泛使用，仍不失为有用的工具。”

2.2　语言距离的计算

客观的语言距离的测量方法是基于语言本身的差

异。Kessler于1995年第一次将编辑距离作为测量爱尔兰方言间的语言距离[16]。从那以后,有很多的研究用这种方法来测量语言或方言间的距离。编辑距离在德国马普所已有实践,获得较好成果。编辑距离被证明测量语言或方言间距离是有效的[17-20]。编辑距离指的是字符串 A 转化为字符串 B 所需的最少编辑数[21]。那么相应地应用到语言学中,一个语言变体的一串语音表达可以相应地对应到另一个语言变体的一串语音表达。编辑距离可以发现一个语音变换为另一个语音所需的最少编辑操作数。假设这反映了语音差异的感知方式和语言演化过程中的变化现象,那么基于任何一个关系词的不同语言的语音表达间的编辑距离,不同语言间的语言距离就可以被计算出来了。

语音字符串之间的距离通过编辑距离算法计算。编辑距离算法可以得到一个字符串变换为另一个字符串所需的插入、删除、替换操作的最小代价,即得到两个字符串之间的编辑距离。该算法的3种操作的“代价”均为1。例如对于斯瓦迪士100词词项“牙”,壮语的发音为[fɐn],傣语的发音为[fan],它们之间的编辑距离为1(ɐ替换为a)。上述计算过程,小的语音差异(如[a]和[a:])跟大的语音差异(如[a]和[ɐ])是等同的,即编辑“代价”均为1。似乎看起来,大的语音差异应该赋予大的距离,但目前并没有语音差异(元音间、辅音间距离的量化)的量化研究。这个问题可以通过将每个元音或辅音符号替换为特征束来解决,每个特征被看作是一个元音或辅音属性,特征束是一系列的特征值,每个值表示对应元音或辅音属性数值化的程度。本文采用Almeida&Braun系统对元音和辅音特征的定义[22]来求得元辅音间距离。这样一来,元音间距离等于元音特征束之间的差异和除以特征数目的平均值。然后这些距离值用于替代编辑距离操作的默认“代价”值1。辅音间距离计算过程类似元音,不再赘述。

通过Python编制程序利用上述思路计算不同语言的两两词汇之间的语音距离。利用词汇距离我们就可以计算语言距离。有了语言距离就可以对语言进行分类了。前文提到,本文利用斯瓦迪士100词进行语言距离计算。所以,当对两个语言进行比较的时候,会得到100个编辑距离。两个语言之间的距离等于100个编辑距离的和除以100。N 个语言之间的所有距离会形成一个 $N\times N$ 的距离矩阵。

3 分类结果

3.1 语言材料

本文收集了16种侗台语族语言的斯瓦迪士100核心词。为方便观察,本节列出了这16种语言的名称及代码。

台语支(壮傣语支):壮语—Zhuang,泰语—Thai,老挝语—Laowo,掸语—Shan,临高语—Lingao。

侗水语支:侗语—Kam(Southern_Dong),佯僚语—Then,莫语—Mak,仫佬语—Mulam,毛南语—Maonan,水语—Standard_Sui。

黎语支:黎语—Hlai。

仡央语支:仡佬语—Gelao,拉基语—Laji,布央语—Buyang,普标语—Pubiao。

3.2 距离矩阵

利用编辑距离算法及上文的计算思路得到侗台语族16种语言之间的编辑距离(百分比表示),如表1所示。

3.3 语言分类树形图

一旦语言间距离计算出来了,有了距离矩阵,那么就可以对语言进行分类了,语言分类结果表明语言之间的关系。本文采用聚类分析技术。随着计算机技术的

表1 侗台语族16种语言之间的编辑距离

	毛南	普标	莫	水	佯僚	布央	仫佬	泰	掸	侗	黎	老挝	壮	拉基	仡佬	临高
毛南																
普标	0.880															
莫	0.363	0.855														
水	0.261	0.863	0.385													
佯僚	0.389	0.885	0.417	0.393												
布央	0.857	0.819	0.859	0.849	0.868											
仫佬	0.440	0.892	0.497	0.428	0.481	0.838										
泰	0.843	0.903	0.826	0.859	0.837	0.903	0.851									
掸	0.840	0.917	0.807	0.850	0.822	0.910	0.845	0.661								
侗	0.437	0.845	0.497	0.412	0.451	0.846	0.477	0.832	0.831							
黎	0.885	0.940	0.881	0.895	0.888	0.967	0.908	0.875	0.925	0.893						
老挝	0.854	0.915	0.839	0.855	0.840	0.944	0.852	0.564	0.680	0.829	0.925					
壮	0.791	0.930	0.811	0.835	0.829	0.907	0.826	0.749	0.766	0.813	0.880	0.781				
拉基	0.958	0.959	0.989	0.992	0.990	0.984	0.979	0.985	0.988	1.004	0.980	0.996	0.943			
仡佬	1.025	0.957	0.995	0.995	0.986	0.909	0.978	0.983	0.990	0.989	0.942	0.961	0.983	0.910		
临高	0.811	0.928	0.827	0.808	0.792	0.917	0.881	0.856	0.846	0.861	0.844	0.844	0.870	0.995	0.932	

发展,聚类分析的技术已经集成到计算机软件中。生物学家开发的一些研究生物种系发生分类的程序,对语言学家很有帮助,因为生物学的分类与语言学分类相类似。聚类分析的结果是一个表示亲缘关系的系统树图,系统树图是一个分层次的结构树,树的叶子节点是不同的语言。Mega是生物信息学上用来构建和绘制进化树的软件,本文利用Mega软件中的邻接法构建语言关系的树状图。

基于表1的侗台语族16种语言之间的编辑距离,生成的语言关系的树状图见图1所示。

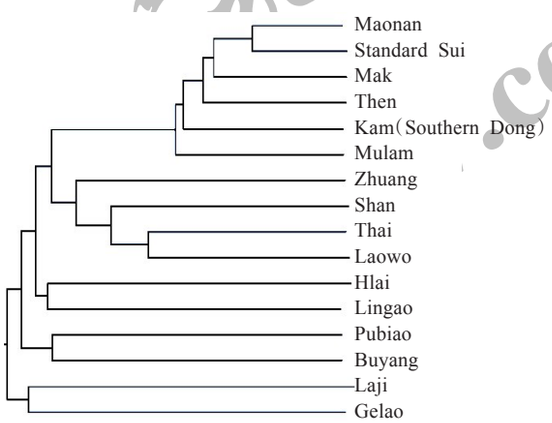

图1　侗台语族16种语言关系树形图

4　分析与讨论

本文的树形图(图1)分类将侗台语族分为5个聚类,即仡佬和拉基、黎和临高、布央和普标、壮傣(壮、掸、泰、老挝)、侗水(毛南、水、佯僙、侗、仫佬、莫语)。树图的第一层为两份,即仡佬和拉基与其他等;第二层为两份,即布央和普标(黎和临高、壮傣、侗水);第三层为两份,即黎和临高(壮傣、侗水);第四层为两份,即壮傣,侗水;第五层为泰和老挝与壮、掸组成一个簇,毛南和水与莫语、佯僙、侗、仫佬组成一个簇。本文分类与前人观点基本一致,尤其是与梁敏的侗台语族谱系树图(见图2)基本

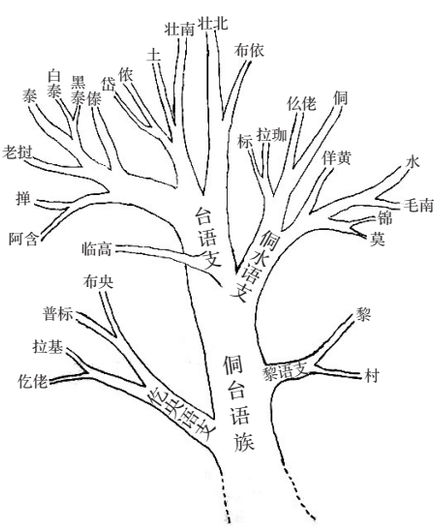

图2　侗台语族谱系树图

相符,比如侗水(毛南、水、莫语、佯僙、侗、仫佬)语支语言之间的关系;壮傣(壮、掸、泰、老挝)语支语言之间的关系;仡佬和拉基之间的关系;布央和普标之间的关系。与国内观点不同的是,本文将黎和临高合并独立一支,黎和临高关系比较近。

本文的树形图显示,仡佬和拉基独立一支,布央和普标独立一支,这与梁敏先生的语言观点是一致的。梁敏先生将仡佬、拉基、布央、普标称之为仡央语群,它们之间的关系如图3所示[5]。

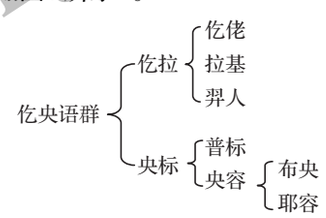

图3　仡央语群

从图3可以看出,仡佬和拉基关系密切,布央和普标关系密切。本文树形结果也是如此。梁敏指出[5]"在仡佬、拉基、布央、普标这几种语言之间,仡佬和拉基的同源词较多……所以仡佬和拉基的关系更密切一些""布央和普标同源的比例也较高,达38.74%,他们的语言系统在仡央语群中也是比较接近的……布央和普标的先民可能有过比较密切的关系和接触来往。""仡佬和拉基比较接近,可以看作是一小团,称为仡拉语支;普标和布央内部也比较一致,又是另外一个小团,可以称为央标语支。"

本文的树形图从整体上来看,相比仡佬和拉基与布央和普标,黎和临高离壮傣和侗水更近,而且从树图树枝的长度来看,壮傣和侗水两个语支中,黎和临高离壮傣语支更近。这与前人研究的观点基本上是一致的。李方桂将台语群分为两个次群:壮语次群和西南次群。他在《中国的语言和方言》[23]中提到"壮语群包括广西大部分地区和贵州南部以及云南东南部所使用的许多方言,使用于海南岛北部临高、澄迈和琼山的熟黎话也属于这个语群。但在海南岛中部和南部的黎话与其他台语相比似乎有很大的分歧。"从李方桂的论述中可以看出,黎和临高与壮语群关系比较近,但是否归属于这个语群还是值得怀疑的。而本文树形图正是将黎和临高单列一支,但又与壮傣语支距离比较近。1987年的《中国语言地图集》和1996年梁敏的侗台语族分类则直接将临高话划归台语支。梁敏用对比研究的方法分析临高语和侗台语族其他语言在语音、词汇和语法方面的异同和它们在发生学上的亲缘关系,认为临高语是侗台语族台语支中的一个独立语言[24]。另外,梁敏[25]指出"临高人的先民属壮泰种族集团的一部分。所以临高语中与台语支相同的语词比它与侗水语支相同的多一些。"邓晓华、王士元认为临高分别与黎和壮的亲缘关系最接近[13]。

与国内观点不同的是,本文将黎和临高合并独立一支,黎和临高关系比较近。国内学者将临高归在壮傣语支,并且认为与壮语关系最近,将黎单独列为一支。本尼迪克特则将黎和临高单列出来。但是也有学者认为黎和临高有关系[13],如法国的萨维那(Savina)认为临高是黎语的一支;德国人类学家史图博(Stubel)提出临高可能是黎语与泰汉语的混合语;邓晓华、王士元认为临高分别与黎和壮的亲缘关系最接近。

5 本文方法与词源统计法的比较

邓晓华、王士元利用词源统计法对苗瑶语族[26]、藏缅语族[27]和壮侗语族[13]做了计量分类。下面从相同点和不同点两个方面将本文方法与词源统计法做对比。

相同点:

(1)二者均借用生物学上关于物种进化关系分析的方法来分析语言的亲缘关系。

(2)二者分析不仅可以显示各种语言的亲疏关系,更可以显示出语言之间的亲缘距离。

不同点:

(1)同源词和借词的问题

词源统计法是建立在同源词统计的基础上的。词源统计分析的基础和前提是核心同源词的选取。如何确定和优选核心同源词是词源统计分析的最重要步骤之一。这个问题一直存在较大争论,同源词有历史文化层次的差别,有的同源词较容易被借用,有的同源词则被借用的概率较低,同源词和借词身份界定很困难。同源词和借词如何区分是个老话题,也是语言系属讨论中最敏感和引起争议的问题,直到现在也未能彻底解决。

本文的编辑距离方法不涉及同源词和借词的问题,无需选取同源词,规避了同源词和借词身份的界定。另外,词源统计法界定的同源词数据不同,得到的结果也会不同,本文方法得到的结果自始至终是一致的,可重复,可验证,科学性较强。

(2)主观性和客观性的问题

词源统计法的操作步骤可简单归结为:编制同源词统计表并计算各对语言的同源比;距离矩阵;绘制树形图。本文方法可简单归结为:搜集各语言的用国际音标标音的斯瓦迪士100词;距离矩阵;绘制树形图。需要说明的,邓晓华在利用词源统计法对语言进行数理分类时,同源词的统计也是用斯瓦迪士100词表(经过了一定的修订,但主体仍然是100词表,个别词有所调整),看各对语言斯瓦迪士100词中有多少词是同源的,从而计算各对语言的同源比。很明显,词源统计法的第一步是经验性的而非理据性的,不同人给出的结果也会不同。确定语言之间同源词的时候仍依赖于专家们的经验和判断,带有主观性成分。

本文直接利用田野调查得到的用国际音标标音的斯瓦迪士100词,后续过程全是计算机自动操作,中间不涉及同源词的选择和确认工作,利用的是语音原生材料,客观性比较强。

(3)全面性和局部性的问题

词源统计法只能做低一层次的语支和语言/方言层级的局部性分析,无法对高一层次的语族和语系层次做出整体的全面性分析。原因是语族和语系层次的同源词选择和确认很难做到,尤其是汉藏语言。

本文的编辑距离方法不仅能做低一层次的语支和语言/方言层级的局部性分析,而且还能对高一层次的语族和语系层次做出整体的全面性分析。这也是本文下一步的研究工作,如侗台语族、汉语族、藏缅语族、南岛语族之间的关系,这也是学界一直有争议的话题,等等。

(4)本文方法特色之处

① 对于新发现语言,可以利用本文方法进行快速分类,再结合历史比较法确定该语言与其他语言之间的关系。

② 本文方法能应用于非常大的语言样本,这有利于大规模语言数据的统计研究和可以揭示之前未知的语言发生关系。

③ 本文方法为长久以来学术界因为传统语言学研究产生的争论提供一种可能的解决方案。

6 总结

本文借助计算机手段,基于斯瓦迪士的100核心词,运用编辑距离算法以及生物学的种系发生树方法,对侗台语族16种语言进行了分类,显示出了侗台语族语言的类簇和分级层次。其结果表明,编辑距离的分类结果与已有的传统语言学的分类结果基本是一致的,其操作过程是可以重复和验证的,可推广至更多的语言及方言的分类,在一定程度上弥补了历史语言学的不足,也为计量法提供了新思路。同时,本文提出了跟传统分类的不同看法,即黎和临高关系非常近,黎和临高合并,在侗台语族中独立一支。另外,本文也进一步验证了斯瓦迪士100核心词可用于语言关系分类的研究中。

参考文献:

[1] 江荻.基本层次范畴与核心词集构建[M].北京:商务印书馆,2008.

[2] 孙宏开,江荻.汉藏语系历史研究沿革[M].南宁:广西民族出版社,2000.

[3] 王育德.中国五大方言的分裂年代的言语年代学的试探[J].语言学材料,1962(8):14-16.

[4] 徐通锵.历史语言学[M].北京:商务印书馆,1991.

[5] 梁敏.仡央语群的系属问题[J].民族语文,1990(6):1-8.

[6] 王士元.A quantitative study of Zhuang-Dong languages[C]//

中国语言学论集,1995.
[7] 毛宗武,李云兵.炯奈语研究[M].北京:中央民族大学出版社,2002.
[8] Brown C H,Holman E W,Wichmann S,et al.Automated classification of the World's languages:A description of the method and preliminary results[J].STUF-Language Typology and Universals,2007,61(4):285-308.
[9] Holman E W,Wichmann S,Brown C H,et al.Explorations in automated lexicostatistics[J].Folia Linguistica,2008,42:331-354.
[10] 陈保亚.论语言接触与语言联盟[M].北京:语文出版社,1996.
[11] 陈保亚.从语言接触看历史比较语言学[J].北京大学学报,2006,43(2):30-34.
[12] 孙宏开,胡增益,黄行.中国的语言[M].北京:商务印书馆,2007.
[13] 邓晓华,王士元.壮侗语族语言的数理分类及其时间深度[J].中国语文,2007(6):536-548.
[14] 江荻.核心词的确切含义及词频导向的构建方法[J].中文学术前沿,2011(1).
[15] 徐丹.研究语言的新视角:语言和基因的平行演变[J].当代语言学,2015(2).
[16] Kessler B.Computational dialectology in Irish Gaelic[C]//Proceedings of the 7th Conference on European Chapter of the Association for Computational Linguistics.Dublin:Morgan Kaufmann,1995:60-66.
[17] Gooskens C,Heeringa W.Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data[J].Language Variation and Change,2004,16(3):189-207.
[18] Gooskens C.The contribution of linguistic factors to the intelligibility of closely related languages[J].Journal of Multilingual and Multicultural Development,2007,28(6):445-467.
[19] Kurschner S,Gooskens C,Bezooijen R.Linguistic determinants of the intelligibility of Swedish words among Danes[J].International Journal of Humanities and Arts Computing,2008,2(1/2):83-100.
[20] Gooskens C.Experimental methods for measuring intelligibility of closely related language varieties[M].Oxford:Oxford University Press,2013:195-213.
[21] Levenshtein V I.Binary codes capable of correcting deletions,insertions and reversals[J].Doklady Akademii Nauk SSSR,1965,163(4):845-848.
[22] Almeida A,Braun A."Richtig" und "falsch" in phonetischer Transkription;Vorschlage zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten[J].Zeitschrift fur Dialektologie und Linguistik,1986(2):158-172.
[23] 李方桂.中国的语言和方言[J].民族译丛,1980(1):1-7.
[24] 梁敏,张均如.临高语研究[M].上海:上海远东出版社,1997.
[25] 梁敏,张均如.侗台语言的系属和有关民族的源流[J].语言研究,2006(4).
[26] 邓晓华,王士元.苗瑶语族语言亲缘关系的计量研究—词源统计分析方法[J].中国语文,2003(3):253-263.
[27] 邓晓华,王士元.藏缅语族语言的数理分类及其分析[J].民族语文,2003(4):1-14.

(上接48页)

[10] 韩丽.社交网络中的信任推荐和好友搜索过滤算法研究[D].秦皇岛:燕山大学,2012.
[11] Zhao H,Wang S,Chen Q,et al.Probabilistic matrix factorization based on similarity propagation and trust propagation for recommendation[C]//Proceedings of Collaboration and Internet Computing,2015:90-98.
[12] 高发展,黄梦醒,张婷婷.综合用户特征及专家信任的协作过滤推荐算法[J].计算机科学,2017,44(2):103-106.
[13] Li G.Pairwise probabilistic matrix factorization for implicit feedback collaborative filtering[C]//Proceedings of International Conference on Security,Pattern Analysis,and Cybernetics,2014:17-25.
[14] 尚志刚,董永慧,李蒙蒙,等.基于偏最小二乘回归的鲁棒性特征选择与分类算法[J].计算机应用,2017,37(3):871-875.
[15] 吴德,刘三阳,梁锦锦.多类文本分类算法GS-SVDD[J].计算机科学,2016,43(8):190-193.
[16] Chag T M,Hsiao W F,Lu C J.Item-level turst-based collaborative filtering for recommender systems[J].Information Management,2007(1).
[17] Xu Q,Zheng S,Cai M.IBCF improved algorithm based on the tag[C]//Proceedings of the International Conference on Computer,Networks and Communication Engineering,2013:265-268.
[18] Kushwaha N,Sun X,Singh B,et al.A Lesson learned from PMF based approach for semantic recommender system[J].Journal of Intelligent Information Systems,2017(4):1-13.