多语种网络空间语言战略研究

[本栏导读] 如何在网络空间传播“中国主张”和“中国声音”,构建网络空间命运共同体,打造全球共通的网络空间话语体系,是习近平同志互联网全球治理体系变革思想的重要任务之一。网络空间语言传播战略已经成为近年来中国语言战略与语言规划研究的新兴领域之一,涉及新时代网络世界的语言活力与语言生态、语言竞争与语言博弈、语言互通和话语建构等核心议题。本栏论文是国家语委重大科研项目“多语种网络空间语言战略研究”(主持人:姜锋)的阶段性研究成果。三篇论文分别从理论、比较与实践角度对网络空间语言生态、语言传播与话语建构议题进行了初步探讨。郭书谏、沈骑基于数据实证研究方法,分析网络空间世界语言使用的活力分布及其成因;杨波从国际比较视野对俄罗斯网络空间语言传播战略进行分析;顾亿青、衣永刚基于高校多语种网站建设实践,探索互联网空间构建人类命运共同体的途径与方法。

互联网空间的世界语言活力及其成因*

郭书谏[1] 沈 骑[2]

(1. 上海外国语大学语言研究院 上海 200083; 2. 同济大学外国语学院 上海 200092)

[摘要] 本文试图通过构建基于经济、人口、内容、媒体和知识五种因素的互联网空间语言活力指标体系,分析世界各种语言的活力指数和排名情况。在此基础上,进一步解释世界语言的活力指数分布及其原因。研究发现:世界各种语言的活力指数呈现不平衡的长尾分布;世界各语言呈现层级化特征。因为随着互联网技术的发展、在线言语社区的形成,个体的语言选择更加倾向于语言层级体系中的上位语言,人口、经济、媒体等资源逐步集中于几种桥梁语言,而大部分地方性语言和弱势语言则在这一资源集中化的过程中失去活力,成为语言活力指数当中的“长尾”。

[收稿日期] 2018 - 07 - 20

[作者简介] 郭书谏,上海外国语大学语言研究院博士生,主要研究语言政策、社会语言学;沈骑,同济大学外国语学院教授,博导,主要研究语言政策与规划。

* 本研究受国家语委重大项目(ZDA135—3)课题资助。

[关键词] 语言活力；网络语言；语言濒危；语言政策
[中图分类号] H002 [文献标识码] A [文章编号] 1003 - 5397(2019)01 - 0027 - 10
DOI:10.16499/j.cnki.1003-5397.2019.01.005

The Vitality and Distribution Reasons of World Languages in the Cyberspace

GUO Shujiang, SHEN Ji

Abstract: This paper proposes a five-factor index system for language vitality evaluation, taking the factors of economy output, population, web content, social media and knowledge into consideration. The index system aims to examine the vitality of world major languages and their rankings and explain the possible reasons for the distribution of the indexes. The findings show that the vitality indexes of world languages are distributed in an unbalanced and long-tailed manner as a hierarchy system. The reasons are that with the development of the internet and the online speech community, people tend to learn and use the language in the higher position of the language hierarchy system. Moreover, the bridge languages have attracted more users, economic resources and media coverage, but most local languages and minor languages are devitalized during the process and show the long-tailed distribution in the index system.

Keywords: language vitality; web language; language endangerment; language policy

引言

随着互联网与人类生活联系日益紧密，网络空间中的语言使用和语言生态逐步成为社会语言学关注的领域。互联网已经成为当今社会信息传播的重要媒介和渠道，如果一国官方语言在互联网中不具备匹配其国家实力的使用人数和内容数量，从长远来看，将对其国际影响力和国家安全造成不利影响。进入 21 世纪以来，网络空间的语言竞争愈发显著，互联网对语言推广和语言传播的重要作用不言而喻。

作为社会语言学和语言政策研究的一个新疆域，互联网中的交际方式、身份构建和信息传播近年来已经成为社会语言学界关注的重要话题。在国内，胡壮麟(2007)较早引入了“计算机中介交流”(Computer-Mediated Communication)的概念并纳入社会语言学分析框架，综述了国外对虚拟言语社区中语言的身份、性别、权势等社会变项的研究。吴东英等(2016)以变异社会语言学和互动社会语言学两种范式为主线，回顾了近十余年国内外主要的新媒体社会语言学研究。国外通过社交媒体等互联网语料进行虚拟言语社区的研究更为丰富。Sharma(2012)考察了尼泊尔大学生们在 Facebook 上的语言使用，探讨社交媒体如何影响年轻人的语言使用。Bamman(2014)采用 Twitter 建立大规模语料库，实证了社交媒体语言中的性别特征。通过以上文献可以发现，社会语言学对于新媒体、互联网的语言研究遵循变异学派的研究范式，更多关注于微观层面。对于语言的宏观分布和相互关系，虽有学者提出相关理念探讨(李

宇明,2016),但仍然具有实证研究的价值和空间。近年来,随着中国积极参与和推进全球治理,特别是在提出全球网络治理新主张的背景下,系统分析世界语言活力,制定提升汉语在网络空间的活力的战略规划,对新时代构建互联网空间人类命运共同体,具有十分显著的作用。

互联网时代的语言活力研究还存在一些基础性问题待解决: 世界各种语言的语言活力如何? 汉语是否处于相对弱势的处境? 语言转用和语言濒危是客观自然的历史过程,还是语言帝国主义等意识形态作用的结果? 本研究以这些问题为导向,试图通过建立语言活力指数的量化指标,探讨互联网中各语种的语言活力问题,揭示世界语言生态中各种语言所处的地位及其背后原因。

一　语言活力的研究概述

语言活力这一概念翻译自英语 language vitality,国外关于语言活力的论述伴随着语言濒危问题而来,二十世纪七八十年代有学者提出了语言死亡(Denison, 1977)和语言转用(Dorian, 1980)。根据语言活力评估语言濒危等级,扭转语言转用成为社会语言学界普遍关注的议题。国内关于"语言活力"的论述多见于民族学和少数民族语言保护研究,黄行(2000)较早提出了语言活力的概念,并通过定量研究方法系统地评估了中国国内少数民族语言在行政、立法、司法、教育、媒体、文艺、出版等十项基本活动中的活力,对各族语言进行了描写测量和分类聚层分析。这一概念后来见于戴庆夏(2006)、陈章太(2008)等关于濒危语言保护的论著之中。

国内语言政策学者研究互联网领域的语言活力问题,常以国家安全作为分析视角。李宇明(2003)最早对互联网领域的语言安全问题做了分析阐述,认为必须"通过语言信息处理和网络数据库的快速建设,争夺虚拟空间的汉语地位"。沈骑(2014)认为"在军事语言安全以外,还应加深和拓展非传统安全领域的语言规划研究",其中互联网是非传统安全领域语言安全的关键领域之一。

互联网的语言活力问题也引起了计算机科学研究者的关注(Paolillo,2005; Pimienta, 2009)。1995 年科托努的法语国家首脑会议期间,有学者提到英语在互联网中的内容占比超过 90%。这一数字不仅引发了业界的普遍关注,也引起了一系列对互联网中各语种内容数量的统计研究。非政府组织网络和发展基金(FUNREDES)最初以当时流行的搜索引擎 AltaVista 为工具,后来也将谷歌纳入其中。1995 ~ 2007 年,这项历时 12 年的研究,利用搜索引擎统计一系列不同语言的同义词汇的出现频数,估算了世界互联网中各种语言的占比情况。另一种针对这一问题的研究来自加拿大的互联网公司 Alis Technologies 和 OCLC 组织,两项研究采用了相似的研究方法: 随机抽取一定数量的网站首页,通过计算机自动识别其所属语言,统计各种语言在样本中的占比,并以此估计互联网总体的语言占比情况,以上两类研究采用的都是抽样方法。随着网络爬虫技术的发展,谷歌率先运用搜索引擎,索引并统计全部互联网页面所属语言的页面数量。

FUNREDES 的研究结果(Paolillo,2005; Pimienta, 2009)表明,以英语书写的互联网页面内容占比从最早统计的 80%(1995)逐步降低至 40% 左右(2007)。2010 年后随着 Facebook、Twitter 等社交媒体的飞速发展,通过搜索引擎的关键词页面数量进行抽样或依据搜索引擎索引的语言占比研究出现了诸多问题,已经无法客观地呈现互联网中的语言生态。首先,不少社交媒体中的内容不再被搜索引擎收录,虽然微信朋友圈每天产生了海量的汉语内容,但由于需

要用户登陆才能访问，无法被百度和谷歌索引收录。其次，随着爬虫技术和人工智能的发展，大量 web 页面内容不是人工创作，而是由机器生成的，不能客观衡量该类语言的活力。再次，仅对语言的 web 页面做统计，未能呈现语言背后所承载的经济、人口等现实因素。最后，不少研究仅对语言的分布做了刻画和描写，未能对语言的分布特征背后的社会和行为做出解释和探讨。

Ronen(2014) 利用维基百科、谷歌图书和推特三大网站中超链接页面的语言数量数据，通过大数据和数据可视化技术绘制了三个网站中语言之间信息传递的全貌。这项研究认为: 以三大网站为例，可以发现世界语言层级化的结构: 英语扮演着互联网信息中心枢纽的角色，连接着德语、法语和西语三个中间枢纽，尽管汉语、印地语、阿拉伯语使用人数众多，但在这三大网站中并不扮演信息枢纽的作用。这项研究创造性地将世界语言之间的关系用定量研究的方法进行验证。

社会语言学领域较早评估语言活力的是 Fishman (1991) 的 GIDS 方法，他根据语言活力的不同将世界语言分为八个级别。后来联合国教科文组织改良并发展出六级濒危评估，绘制了世界语言的濒危地图，Lewis 等(2010) 进一步拓展了这一评估方法。联合国教科文组织依据六要素对语言活力进行评估: 代际语言传承、使用一种语言的绝对人口、总人口中使用该语言的比例、现存语言使用的领域、新领域和媒体的反馈以及语言教育材料的数量; 每一个要素划分为五个级别。联合国教科文组织认为“全世界 97% 的人口使用了大约 4% 的语言，约 90% 的语言将在 21 世纪被代替”(UNESCO，2003) 。社会语言学采用的方法大多以质性描述为基础，进行抽样后对样本进行质性评估，根据样本是否满足每个级别和矩阵中的关键性问题来判断该语言是否濒危。

二 研究设计

本研究假设现实世界的语言活力仅与某种语言的经济和人口两个因素有关。一种语言的使用人口是语言活力的生物基础，人口是语言使用的载体，语言衰亡的直接表现和最重要原因就是原有人口不再使用该语言。某种语言使用者所产生的经济活动总和可以通过经济产出进行核算，一种语言哪怕被较少人口使用，只要使用这一语言的人口族群经济产量较高，那么将具有更多的经济和物质优势让语言得以传承。人类在进入互联网时代之后，虚拟世界的语言使用也成为衡量语言活力的一个重要领域，虚拟世界的人类活动可以三类活动作为代表进行抽样: 网络社交(以推特为代表的社交媒体) 、信息获取(使用搜索引擎) 和知识传播(以维基百科为代表的知识库网站) 。

研究中所采用的各项数据来源于国外两大研究机构，其中维基百科、推特、GDP、人口四项数据来源于麻省理工学院媒体实验室[①]，不同语言的内容占比来源于互联网研究机构 W3Tchs[②]，以此构建了五个指标的量化评价系统。在原始数据的基础上，进行了两步数据整理工作。首先将原始数据标准化，以各个指标中英语的数据作为分母 1，将其余语种数据百分比化。然后对五个指标的百分比数据进行加权平均，权重设计参照表 1。语言活力指数的计算公式如下:

$$Q_i = \sum \frac{a \mid X_i}{X_{eng}}$$

其中 Q 为语言的活力指数，a 为指标权重，X_i 为某个语种某个具体指标的原始数据，X_{eng}

为该指标对应的英语数据。

表 1　互联网时代语言活力评估量化指标体系

评估要素		设计权重(%)	基础指标
现实世界指标	经济	30	语言的 GDP 产值
	人口	30	语言的使用者人口
虚拟世界指标	内容	20	语言在互联网的内容占比
	社交媒体	10	语言在推特上的人数和条目数
	知识传播	10	语言在维基百科上的编辑者和条目数

通过计算,将结果根据语言活力指数由大到小排列,结果见图 1。

语言	语言代码	经济指数	人口指数	内容指数	推特指数	维基指数	语言活力指数
English	eng	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Chinese	zho	29.34%	105.00%	3.91%	0.18%	3.68%	41.47%
Spanish	spa	16.97%	33.33%	9.96%	17.31%	6.88%	19.50%
Arabic	ara	9.35%	35.33%	1.37%	3.91%	1.10%	14.18%
Japanese	jpn	9.28%	8.80%	11.13%	35.90%	8.14%	12.05%
German	deu	14.32%	12.33%	10.94%	0.67%	17.13%	11.96%
Russian	rus	8.71%	18.53%	13.09%	1.79%	6.27%	11.60%
Portuguese	por	6.95%	19.33%	5.08%	18.22%	2.61%	10.98%
Malay	msa	3.66%	20.00%		19.40%	0.49%	9.09%
French	fra	6.11%	13.33%	8.01%	1.34%	11.63%	8.73%
Italian	ita	4.34%	4.67%	4.69%	0.62%	6.01%	4.30%
Korean	kor	3.43%	5.20%	1.76%	4.57%	1.33%	3.53%
Persian	fas	2.13%	7.13%	3.32%	0.03%	0.81%	3.53%
Turkish	tur	2.07%	4.67%	2.73%	1.83%	1.04%	2.85%
Dutch	nld	2.21%	1.80%	2.54%	4.12%	3.22%	2.45%
Polish	pol	1.77%	2.87%	3.32%	0.07%	3.32%	2.39%
Thai	tha	1.39%	4.87%	0.59%	2.92%	0.46%	2.33%
Vietnamese	vie	0.56%	5.33%	1.17%	0.06%	0.58%	2.07%
Ukrainian	ukr	0.66%	3.00%	0.39%	0.01%	0.93%	1.27%
Romanian	ron	0.64%	1.87%	0.98%	0.03%	0.43%	0.99%
Czech	ces	0.66%	0.80%	1.76%	0.04%	0.86%	0.88%
Swedish	swe	0.81%	0.67%	0.98%	0.23%	1.78%	0.84%
Hungarian	hun	0.57%	1.00%	0.98%	0.04%	1.37%	0.81%
Hebrew	heb	0.63%	0.67%	0.39%	0.03%	2.76%	0.74%
Serbo-Croatian	hbs	0.60%	1.53%		0.02%	1.02%	0.74%
Modern Greek	ell	0.81%	1.00%		0.21%	0.36%	0.60%
Finnish	fin	0.44%	0.40%	0.59%	0.02%	1.48%	0.52%
Bulgarian	bul	0.33%	0.80%	0.39%	0.01%	0.57%	0.47%
Catalan	cat	0.56%	0.60%		0.09%	0.78%	0.43%
Danish	dan	0.45%	0.40%	0.59%	0.03%	0.49%	0.42%
Norwegian	nor	0.54%	0.33%	0.20%	0.07%	0.90%	0.40%
Slovak	slk	0.33%	0.47%	0.59%	0.01%	0.22%	0.38%
Lithuanian	lit	0.15%	0.27%	0.20%	0.00%	0.18%	0.18%
Slovenian	slv	0.12%	0.13%	0.20%	0.01%	0.23%	0.14%
Estonian	est	0.04%	0.07%	0.20%	0.01%	0.18%	0.09%

图 1　语言活力指数排名③

图 1 显示,英语、汉语、西班牙语、阿拉伯语、日语、德语、俄罗斯语、葡萄牙语、马来语、法语分列前十位。英语活力指数比第二名到第六名的语言活力指数加总更高,处于绝对强势地位。排名前十种语言加总的语言活力指数占语言活力指数总值的 87.77%,显示出几种关键语言占据着世界上绝大部分经济、人口和互联网内容资源。

汉语虽然距离英语的霸权地位仍然有不小差距,但是与其他语言相比更为强势。这得益于中国丰富的人口资源和经济的快速发展,汉语的语言活力在人口和经济两项指标上占据领

先位置。互联网内容占比指数中，汉语由于数据原因受到低估，因为网络访问的限制，国外搜索引擎很难全面索引到国内网络上的内容。仅根据 Alexa 网站流量排行，全球前 20 名流量的网站中，来自中国的网站有 7 个，其中所使用的语言绝大部分是中文。汉语网络世界中存在功能类似维基和推特的平台，形成了较为独立的信息平台，使汉语使用者与其他语言的使用者对话和联系相对较少。

此外，作为拉丁美洲的地区通用语，西班牙语仍然位居世界语言活力前列；阿拉伯语是中东地区重要的通用语言、伊斯兰文化的传承语言，语言活力非常显著。殖民时期的强势语言如法语、荷兰语等语种虽然历史上曾广泛传播，但活力指数不及日语和德语。泰语、马来语两个亚洲语言凭借人口优势在社交媒体上占优势。希伯来语作为一个近百年来伴随着犹太复国运动“复活”的语言，语言活力超过许多具有悠久历史的语言。就语言的发源地而言，活力指数较高的语言都发源于欧亚大陆，目前，发源于非洲的斯瓦西里语、刚果语，以及曾经通行于美洲的印第安土著语言和新西兰的毛利语都难觅其踪迹。

将语言活力指数由大到小排列绘制柱状图，并绘制累计曲线，如图 2 所示。

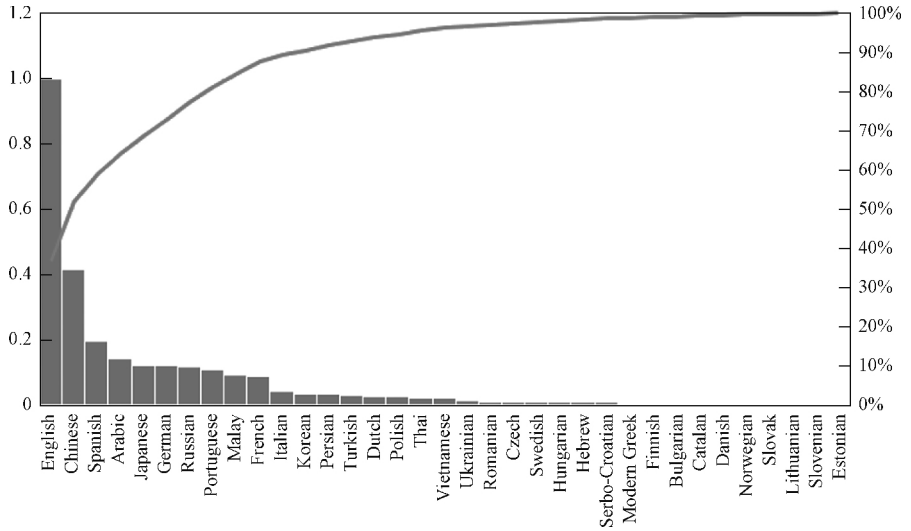

图 2 各语种的语言活力指数分布

通过观察发现：语言活力指数的分布呈现“长尾分布”的特征：即大部分语言指数分布在较少的几种语言中，其他大部分语言的活力指数非常小。从构成要素可知，语言活力指数反映了某种语言的经济、人口、媒体、知识等要素。这种长尾分布显示出世界各个语言之间活力存在极大的不平衡，少数几种强势语言占据了大部分人口、经济和互联网资源，而大多数语言处于弱势地位，未来恐将进一步失去活力。

三 语言活力分布的成因

我们认为语言活力指数呈现长尾分布主要有以下两方面原因：一是世界语言在互联网信息传播中的地位不均，具有层级化特征；二是受社会资本的影响，个体的语言选择倾向于学习和使用上位语言。

Kelly-Holmes（2006）分析了 548 个互联网经贸网站，Ronen（2014）统计了谷歌图书翻译的

语言联系,二人提出了一个相似的观点。随着互联网的不断发展,母语为不同语言的族群之间交流日益增加,世界语言生态开始形成层级化的特征: 即英语开始扮演互联网通用语的角色,而汉语、西班牙语、阿拉伯语等则扮演着联系各个地方语言的功能。依靠为数不多的十来种关键语言,构成了虚拟世界中不同语言之间的信息流动。假使某项信息由位于巴塞罗那的居民用加泰罗尼亚语书写而成,一个粤语使用者想获取该信息,大致可能的路径是: 加泰罗尼亚语—西班牙语—英语—汉语普通话—粤语。换言之,在语言网络中位于较低层级的语言之间缺乏直接的联系,几乎没有使用者能够同时使用粤语和加泰罗尼亚语,而是通过网络中的节点语言实现人际传播。作为个体,语言使用者更倾向于使用语言层级网络中的上位语言,因此人口、经济等要素逐步向少数几种语言集中,而大量的地区语言和小众语言则逐步失去活力,成为语言活力指数分布中的"长尾"。

这种语言层级化的社会现象并非自古以来就是这样,而是随着互联网的发展、不同母语族群之间交流的加深而产生的新现象。人类早期由于地理分隔交流较少,各个语言之间独立发展,相互联系并不紧密。随着军事扩张、经贸往来的增多,不同语言的人群开始接触并相互沟通,形成了若干种地区通用语。特别是新航路开辟之后,英语、法语、西班牙语被殖民者带到了被殖民地国家,成为当地的重要语言,不同语言之间的人口、经济等社会资源逐步产生差距。英语在北美洲、法语在非洲、西班牙语在拉丁美洲逐步成为了社会精英阶层和殖民者的语言(社会语言学中称为高变体) ,而当地语言仍然作为平民阶层的语言(社会语言学称之为低变体) ,在诸多殖民地国家形成了独具特色的双言现象。随着 20 世纪 90 年代互联网的兴起,语言交流频率增加、门槛降低,英语作为互联网的主流语言,在信息时代获得了空前的强势地位,逐步成长为连接法语、西班牙语、阿拉伯语等地区通用语的全球通用语言。

面对语言在信息世界中的层级分布特征,个体更加倾向于学习、掌握和使用上位语言,以期获得更多社会资本,在竞争中获得优势。在世界语言层级体系中,下位语言通常被用来在某一族群内部进行沟通,而不同族群之间相互沟通时则使用上位语言。中国人之间的网络交际通常使用汉语,而中国人与美国人交流采用英语。以中国的社会现实为例,父母一代的人群能够掌握地方方言和普通话,而子女一代较为普遍地掌握英语和普通话。以互联网使用场景之一的微信为例,父母一代之间通常会用当地方言进行语言交流,而子女一代则通常使用普通话,当子女同其他国家和民族的人交流时使用英语,如图 3 所示。这一现象也得到国内微观层面家庭语言规划的证实,张治国等(2018) 通过实证调查发现汉语方言的使用将随着代际的出现而呈下降趋势,而外语(英语) 的发展趋势正好与此相反。这一趋势使得越来越多的资源和人口向强势语言集中,而弱势语言则进一步失去活力。

语言的层级式分布和个人的语言选择背后具有深刻的社会因素和利己动机。子女一代通常不再学习或者使用方言,而更多地学习和使用英语。这一现象很大程度上取决于语言工具主义: 视语言为工具,希望通过学习掌握并使用某种上位语言以获得更好的工作机会和社会地位。学习并使用位于语言网络中较高层级的语言,对于个人而言具有重要的信息、社交和文化象征意义。出于个人发展的利己动机,在掌握一门语言即可以满足生存基本需要的前提下,掌握一门更高层级的语言可以在信息、社交、升学、就业等方面取得优势地位。

这种微观层面个人从利己的角度出发所做的语言选择,宏观上促进全世界范围内的信息交流更加便捷通常。例如: 如果 5 种不同语言的人意图实现相互理解沟通无碍,那么任何个体

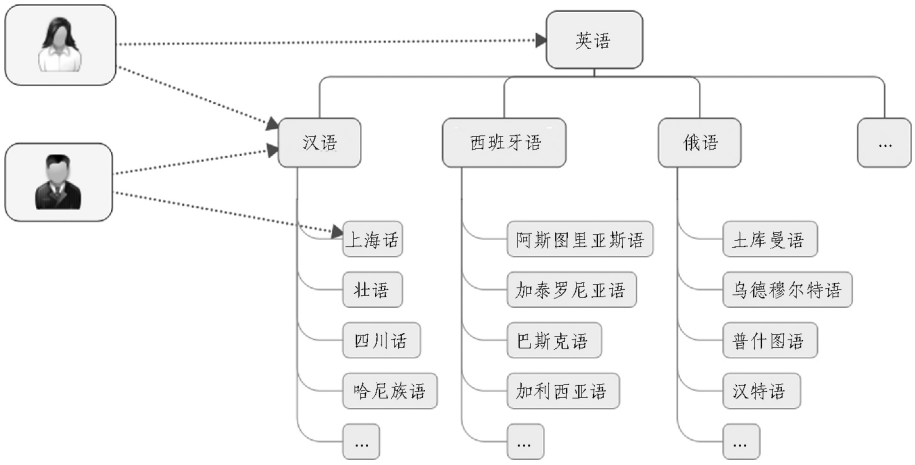

图 3 语言层级网络中的双言现象④

都需要掌握 4 种外语，总计构成 10 种语言组合。但是如果以 5 种语言中的某一种语言作为桥梁语言，任何个体只需要掌握本族语和通用语，即可实现和其余语言使用者交流，共计 4 个语言组合。无论从学习者的困难程度，还是从经济成本考量，第二种方式无疑更加具有效率和优势，更有利于世界各个民族之间的交流和了解。如图 4 所示。

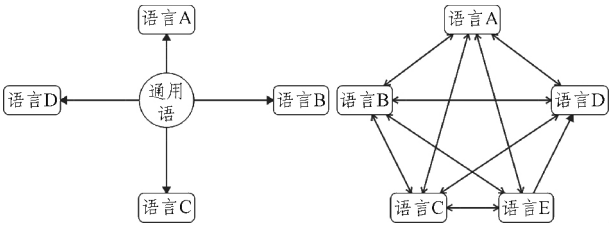

图 4 通用语在世界语言体系中的作用

以上例子仅含有 5 种语言，然而世界范围内使用人口超过 5000 万的语言就有 13 种之多。在经济全球化背景下，为了信息能够更有效传播，必然会形成某种充当信息桥梁的通用语。这种趋势并非某个国家政府或者机构刻意为之，而是适应社会经济发展需要的必然产物。因此，从这个意义上讲，作为人类信息交流的工具，语言多样性减少，仅有少数几种语言具有较高的语言活力是客观的、自然的发展结果，并非如 Phillipson（1997）所认为的语言帝国主义⑤。语言多样性并非天然的，而是历史发展阶段的产物。伴随着科技进步和全球化的进一步发展，原有的地理疆界的阻隔被互联网技术拉近。每个公民，作为个体怀着利己的动机选择学习和使用语言，具有较高社会资本和人力资本价值的语言通常在个体选择中获胜，扩大了使用人口，占据了更多的经济资源。尽管宏观上我们看到世界语言多样性在逐步减少，语言转用现象在广泛发生；但从微观上讲，这是由千万个微观个体在全球化时代、互联网时代的个体选择共同引起的。近年来，不少语言政策学者开始反思宏观语言政策为什么经常失败（Kaplan，2011），其根本原因在于每个社会个体、每个家庭从微观层面对语言政策具有自主决定权和选择权。尽管宏观政策文本能够规定、规范和指导，但最终的语言选择仍然得由个体做出。互联网时代，随着在线言语社区的产生，不同语言族群打破地理的阻隔交流日益加深，强势语言无论在社会资本还是交际功能方面都居于强势地位，个体的语言转用更加容易发生，加快了弱势语言濒危的速度。因此从濒危语言保护的角度而言，互联网虽然为濒危语言提供了数据化保存的机会，但由于个体在网络生活中减少了弱势语言的使用，长期而言互联网的发展将加速弱势语言的濒危化。

四　结语

语言活力是一个社会语言学长期以来关注的话题,随着互联网时代的到来,语言活力的研究也引起了计算机科学的关注。本研究试图通过构建经济、人口、内容、社交媒体、知识传播为基础的世界语言活力指标体系,量化评估世界范围内各语言的活力。这一量化指标体系通过各语言的 GDP、人口、Web 内容占比、推特条目、维基条目的加权平均结果,计算世界主要语言的活力指数。值得注意的是,语言活力的量化评估仍然是抽样统计,指标体系的设计和统计方法的选择对结果都将产生一定影响,如何更具代表性、更客观地进行分析,避免统计误差,是研究过程必须重点关注的问题。未来语言活力的量化评估体系仍然具有进一步拓展多因素的空间,语言的区域覆盖范围、国际经贸、学术发表等因素均可通过相关研究数据纳入指标体系之中并赋予相关权重。此外,本研究所选取的数据源对于中文互联网世界的语言活力刻画稍显不足,未来国内互联网领域的相关统计研究存在广阔空间,定量研究和统计学将为语言活力研究拓展新的空间和疆域。

研究结果显示世界各主要语言活力呈现长尾分布,强势语言和弱势语言的语言活力不均衡现象显著。随着技术的进步和全球化的发展,互联网时代的世界各语言之间呈现出层级化的特征,由于个体在语言选择和语言使用中倾向于使用上位语言,人口、经济、媒体等资源语言活力的构成要素逐步朝强势的桥梁语言集中,大量少数族裔语言逐步失去活力,成为语言活力指数分布中的"长尾现象"。个体出于自利目的的语言选择,客观上进一步强化了层级网络体系中强势语言的地位。互联网时代的语言层级化和语言活力的"长尾现象"是全球化和技术发展的必然结果,根源在于微观个体更加倾向于使用信息时代语言层级中的桥梁语言,从而引发人口、经济、媒体等资源向较少的语言集中,而大量地方语言和弱势语言逐步失去活力。在这一宏观趋势之下,语言保护不仅要注重语言的建库保存(曹志耘,2015),也要重视提高少数族裔语言的社会资本价值。在全球化和互联网技术发展的时代背景之下,资源向几种强势语言集中、语言转用不断发生、各个语言之间逐步层级化、信息中枢语言的出现等现象是社会发展的历史进程,其微观基础在于个人的利己选择和语言工具主义的影响,这一趋势客观上促进了语言作为信息传播和交际工具的经济效率。

[附　注]

① 参见 http://language.media.mit.edu/data。

② 参见 https://w3techs.com/technologies/history_overview/content_language。

③ 根据 W3techs 数据,列表中个别语言在互联网 Web 页面中内容太少,未能进入统计排名,因此留空。

④ 所示语言层级仅代表互联网中的语言信息传播关系,语言区中的地方语言与桥梁语言之间存在密切的信息传播关系。如汉语区中与之相连接的有上海话、壮语、四川话等,这些语言之间存在信息传播关系,而汉语则是这一系列语言的信息中枢和桥梁。图 3 所揭示的语言层级不代表语言地位和法定关系。

⑤ 语言帝国主义:强势语言的国家和政府利用经济、科技、信息等多方面优势,促使不发达国家和地区的人口转用强势语言。

[参考文献]

[1] 陈章太. 论语言资源[J]. 语言文字应用,2008,(1).

[2] 曹志耘. 中国语言资源保护工程的定位、目标与任务[J]. 语言文字应用,2015,(4).

[3] 戴庆厦，张景霓. 濒危语言与衰变语言——毛南语语言活力的类型分析 [J]. 中央民族大学学报: 哲学社会科学版，2006，(1).
[4] 黄 行. 中国少数民族语言活力研究 [M]. 中央民族大学出版社，2000.
[5] 胡壮麟. 计算机中介交流的社会语言学思考 [J]. 外语电化教学，2007，(1).
[6] 李宇明. 信息时代的中国语言问题 [J]. 语言文字应用，2003，(1).
[7] 李宇明. 语言竞争试说 [J]. 外语教学与研究，2016，(2).
[8] 沈 骑. 非传统安全领域的语言规划研究: 问题与框架 [J]. 语言教学与研究，2014，(5).
[9] 吴东英，李朝渊，冯捷蕴. 新媒体的社会语言学研究: 回顾与展望 [J]. 当代语言学，2016，(4).
[10] 张治国，邵蒙蒙. 家庭语言政策调查研究——以山东济宁为例 [J]. 语言文字应用，2018，(1).
[11] Bamman D，Eisenstein J，Schnoebelen T. Gender identity and lexical variation in social media [J]. *Journal of Sociolinguistics*，2014，18，(2).
[12] Denison N. Language Death or Language Suicide? [J]. *International Journal of the Sociology of Language*，1977，(12).
[13] Dorian N C. Language Shift in Community and Individual: The Phenomenon of the Laggard Semi-Speaker: International Journal of the Sociology of Language [J]. *International Journal of the Sociology of Language*，1980，(25).
[14] Fishman J A. *Reversing Language Shift* [M]. Multilingual Matters Ltd，1991.
[15] Kaplan R，Baldauf R，Kamwangamalu N. Why educational language plans sometimes fail [J]. *Current Issues in Language Planning*，2011，12，(2).
[16] Kelly-Holmes H. Multilingualism and commercial language practices on the Internet [J]. *Journal of Sociolinguistics*，2006，10，(4).
[17] Lewis M P，Simons G F. Assessing Endangerment: Expanding Fishman´s GIDS [J]. *Revue Roumaine De Linguistique*，2010，55，(2).
[18] Paolillo J，Pimienta D，Prado D. Measuring Linguistic Diversity on the Internet [J]. *Language Magazine*，2005，(7).
[19] Phillipson R. Linguistic Imperialism [M]. *Oxford University Press*，1992.
[20] Pimienta D，Prado D，Blanco á. Twelve years of measuring linguistic diversity in the Internet: Balance and perspectives [R]. 2009.
[21] Ronen S，Goncalve B，Hu K Z，et al. Links that speak: the global language network and its association with global fame [J]. *Proceedings of the National Academy of Sciences of the United States of America*，2014，111(52).
[22] Sharma B K. Beyond social networking: Performing global Englishes in Facebook by college youth in Nepal 1 [J]. *Journal of Sociolinguistics*，2012，16(4).
[23] UNESCO. Language Vitality and Endangerment [R]. 2003.