民间文化论坛

FOLK CULTURE FORUM

2014年第05期

中国民间文艺家协会主办

民间文化论坛 双月刊 2014年第5期（总第228期）

故事计数：论计算方法在民间叙事研究中的应用*

[美]约翰·劳顿（John Laudun）著　宋颖译

[摘要] 本文尝试结合计算机互联网技术和语法或形态学，用一种可能的民俗学方法，来探讨更大领域中正在进行的对人类文本和意义制造的研究。借助新的计算机技术对民间叙事中的语词进行计算和统计，不仅有助于我们简单探讨文本与文本性，了解构成一则文本需用语词的数量及其属性，了解人们如何用语词来创造世界，而且将帮助我们进一步发现那些在传统研究方法下难以发现的文本之间的关系，同时，通过对民俗文本新的观照并探寻我们以前没有注意到的文本间关系，将会引领我们走向新的知识形式，这些形式将重新定义并拓展当前获取知识的途径。

[关键词] 计算方法；民间叙事；路易斯安那宝藏传说
[中图分类号] C953；G122　　[文献标识码] A　　[文章编号] 1008-7214（2014）05-0020-16

我想先谈谈我们面对的不太熟悉的新主题：互联网搜索引擎的空框子。当今时代，我们都不会对这样的现象感到惊讶：当我们敲击键盘时，这个空框子就会开始预测我们搜索的可能完形。例如，当我敲出“NATU”，搜索引擎为我提供了“Nature，Natural selection，nature's sunshine，natural family planning”等语词。当我输入完搜寻的第一个词，实际上这个词是“Natural”，我开始敲入要搜索的第二个词的开头“L”，搜索引擎又提供了“natural log，natural law，and natural laxatives”等语词。当我继续敲击出字母“A”，引擎会正确地猜出我要输入的词是“Natural Language”，这正是我有兴趣的主题。（参见图 1）

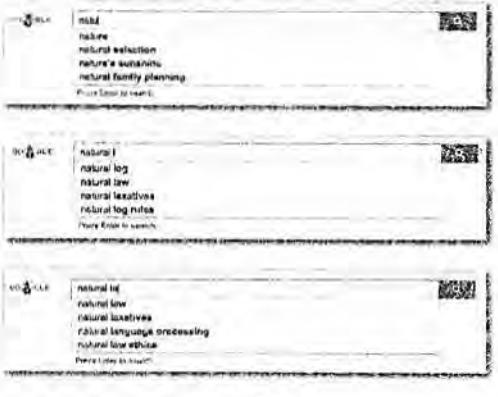

图 1

当然，搜索引擎在我录完前就能完成搜索结果的能力，是自然语言处理(NLP)的编入功能进入了搜索的过程。

[作者简介] 约翰·劳顿（John Laudun），美国路易斯安那大学英语系副教授。
[译者简介] 宋颖，中国社会科学院助理研究员。

* 这篇论文是我 2013 年 8 月应邀在中国社会科学院民族文学研究所演讲的扩充版本。感谢朝戈金所长对我的邀请，以及他和民族文学研究所对我的接待，感谢赵宗福院长邀请我参加青海社科院组织的有关昆仑文化研究的国际会议。在北京和青海停留期间，我从不少同行的观点和言谈中深受启发，尤其是要感谢中国社科院的安德明、北京师范大学的杨利慧和印第安纳大学民俗学研究所的苏独玉等教授，正是由于他们的支持和鼓励，本文在对话中得到了进一步的完善。当然，文中倘有任何不足，都是因我个人力所不逮而造成。

本文系作者应《民间文化论坛》之约，特为本刊撰写，谨此致谢！——编者注。

作为诸多交叉研究领域、特别是语言学和计算机科学范畴的一项重要成果，自然语言处理技术在过去的十年间已经取得了重大的飞跃，这归功于来自世界各地数以百万计的用户提供了大量的自然语言材料，他们利用交流机会借助互联网这种透明的平台，使用着他们选择的语言——自然的语言（自然而然地）。从古德堡项目中成千上万的小说，到数以百万计的推特和脸书每日的更新，我们慢慢进入到这样一种阶段：人类手写表达的全部内容都可以用计算机分析、建模和程序化。

当然，手写往往受限于经由人类手指敲击键盘而直接输入的内容，以及扫描和光学符号识别可能造成的错误。任何用过光学符号识别（OCR）的人都会告诉你，自动化输入长篇打印文本或很多此类文本的过程，是指大量细致地创建、管理、以及沉闷的数小时手工纠错等等。手工纠错这种输入过程，是具有完整的架构而将人类置入其中的过程，谷歌的 Captchas 项目[①] 就为交互的用户提供了从谷歌图书的篇章到精心制作的各种历史档案，将学者也卷入了证明的过程。

但是口头话语呢？大多数人都觉得“自然的”语言能够大量在线找到，这些对话出现在小说戏剧、法庭实录、以及网络上公开发表的手稿或口头故事中。（因此，在某种程度上，更多地制作这类材料并使其符合开放存取组织的标准，正是民俗学者的诉求。）

那些细致转写的能够表现出传统的说话方式或语词，对我们世界的创造性的口头材料相对缺乏，然而，这并不能令民俗资料家或民俗理论者扫兴。就此而言，口头话语在吊信息科学家的胃口，原因如下：首先，比起大量的书写话语而言，口头话语更像是很多人所写的，正如斯蒂文·伯德、伊万·科雷、爱德华·洛佩尔的 Python 语言（是一种面向对象、解释型计算机程序设计语言）著作中介绍自然语言处理时提到的：

> 这需要技巧、知识和一点运气来解开此类问题的答案，正如：我预算有限时在费城宾夕法尼亚或匹兹堡之间该去哪处落脚呢？专家们讨论数码单反相机到底在说什么？上周靠谱的评论员是如何做出钢铁市场的预测的？让计算机来回答这些问题自动地涉及一系列语言处理任务，包括信息提取、推理和摘要，还需要算出数值范围，具有一定的鲁棒性[②] 水平，我们现在还不能实现。（*Natural Language Processing with Python*，p27）

而且，正如上引这段话中所列问题的这种交流属性所体现的，语音识别，显然是研究的前沿。也许一些人用过苹果语音助手（Siri）或者谷歌即时（Google Now），来为列表增加一个提醒，或寻找一个人或一个地方。那些使用语音识别软件工作或与基于语音识别的语音助手交互的人，已经明显意识到其局限性了。本地端的语音识别软件常常需要大量的互动训练：用户要训练软件来辨识特定的语型变化，软件告诉用户它能识别或不能识别什么——不能识别的就无法智能地说出、理解和转化。像语音助手的服务是把说出的话语代传给庞大的服务器群，并应用超量的计算能力来简单地解析某种表达，如“把牛奶加入我的食品杂货清单中”。

除了口头话语中言说这个维度之外，信息学家们还对各种民俗文类很感兴趣。在宏观分析层面上，特定的文类长期稳定而广泛的出现已经吸引了网络理论家们研究“复杂系统的性能如何在内容之间的相互交流中凸显”（Carron 2012:1）[③]。在微量分析层面上，研究者已经着手研究，如何对传统的话语进行可靠的解析，就像现在用大量自然语言处理算法来处理句子一样，后者事实上已经转向弗拉基米尔·普

---

① 图灵测试，是一种全自动化公开的挑战，即通过回应来判断用户是计算机还是人类。——译者注。

② 鲁棒性：自动化术语，原始载体在经历各种信号处理过程后，隐藏信息仍能保持完整性或仍能被准确鉴别，不因处理后而导致信息丢失。——译者注。

③ 这里要注意的是，在一定程度上，一种假设认为口头话语是“自然”世界以某种其他话语形式所不具备的方式而进行的拓展。我还没有看到科学家和人类学家在这个问题上有过稳健的讨论。

罗洛普的工作，在探讨他对一组俄罗斯民间故事的有影响力的形态学分析如何应用到其他叙事材料的可能性。普罗洛普的《故事形态学》对于民俗学者来说都很熟悉。

人们敬佩他的精确性，但是一般却不知道还能从哪些方面去思考文本或对文本还有什么处理方式，类型研究再现了以前借助文类将文本目录化的研究取向，而不是去了解那些常常会再现社会现实的文本是如何被用于社会现实的创造的。对于许多人来说，那些有关19世纪中叶到20世纪中叶一段时期的较大的文本集成，属于残留的概略性数据，它是当时学术研究更注重于跟踪思想进展和人类历史的时代产物。

但是对这些大规模的项目还有另外一种维度的研究，当然，似乎应该说是计算机学家偶然发现了它：即为人类文本的制造而创建出一个具体的工作模型、某种可供识别的叙事模型的可能性。马克·芬雷森（Mark Finlayson）指出，他迷上民间故事是因为它们“具有达尔文式的自然选择过程，叙述的每一个部分都与所承载或详述的文化是和谐一致的，那些不一致的就会被歪曲或摒弃”（2009：127）。芬雷森研究的目的，如其他人的目的一样，是要发展计算分析方法来研究一组给定的故事，还要加速处理过程，同时去除那些因很多人为影响而形成的含糊歧义。也就是说，例如，一种特殊形态中有多少因素是受分析者的影响而形成的，后者可能已受到其他形态的影响；这种形态在多大程度上影响了被考察的叙述；与这种形态的确认相关的内容必须包含哪些因素，等等。很多方法，受研究者的兴趣在推动着，像贝叶斯算法模型和相似故事合并法，可惜都是基于将叙述变为计算机可读形式的转化代码。通常这种转码过程意味着大量大块的话语被删减为情节的概述，分析者必须、也必然地，将实际的语词转化成适用于分析的术语，即能够将一种叙述与另一种叙述相比较。芬雷森的工作很引人注目，也是基于这种方式。

结果是科学家的研究常常止步于以计算机语言模型来建立某种方法，这些方法已经证实是要高度依靠科学的水准，然后才能应用于更大范围、更复杂程度的集合，如文本集。再回到芬雷森的例子，他开始研究少量莎士比亚戏剧的句子来验证他的方法论有效，这样一来，他就可以从十个词的文本拓展至上千词的文本。

当然，对于文本的研究还有中间地带，这个中间地带对于民俗学者来说也很熟悉，它包括全世界的人们用以安排每一天并维持他们置身其间的现实生活的大量文本。这样的文本可以是短小的谚语，或者是长一点的个人故事，或者是他们能引述的长篇史诗或纪事等。感谢近半个世纪的民族志方法，我们了解到这些文本不仅有效，还有相当精准的记录，关于它们现行的地域，以及在社区内传承的演唱者等等。因此，我们有大范围的文本，从几个词到上万词的，相当自然地形成的文本。在很多个案中，此类文本是来自单一的社区，但是也有很少量出现在可敬的考古学记录中的人类话语产物，我们可以以此为上述正在进行中的探究贡献新的思路。

这篇论文要用一种可能的民俗学方法，来探讨更大领域中正在进行的对人类文本和意义制造的研究。我的这一工作的目标是，探讨观念和故事中的事件之间动态的交叉点：观念即组合起来能够建构出像主题或意识形态的思想；故事中的事件，我们一般理解为连续性的、结构化的。一手是网络或互联网，一手是语法或形态学。项目当前所关注的主要是传说，因为对于我来说，它们提供了多种有利条件可进行探讨：它们通常是最具梗概的故事；通常形式很灵活，即便前后顺序颠倒还保有多种稳定的意义；它们常会自然地连接，卷入很多会说话的主体，使人相信意义确实是共有的。我想澄清的是，我刚开始进行叙事的计算研究，我想大量的口头文类有某种有趣的中间地带，通过表演当中的重复已经有力地证明是具有意义的话语，同时还是非常有用的居中标准（我们敢说是“中央”吗？）当然，这些材料都是民俗研究的对象。

路易斯安那州宝藏传说的研究自成一种分析圈：尝试进行小规模的口头故事集成的形态学研究，我们先开始研究相当概括的摘要，然后将摘要移到提升结果的图层集合中，以致于所有的传说都具有同一种形式。接着我们提升研究的细粒度，想发现有趣的差异，这有可能引出更好的问题。当我们开始研究

一大块内容时，最终还能得到词语——分析关于人类实际上在这个世界中是怎样并如何做事情，我们更感兴趣的是用来组建的团块材料。事实上，正如提姆·唐克里尼（Tim Tangherlini）在他的研究中所指出的，计算的民俗研究应该是什么样，他用了宏观的说法来描述，其中最重要的是，这样的方法确保了民俗文本中的表演维度变成了我们研究中的关键："将人们与地点和时间相连，使得文集凭借此基础，为民俗复杂多变的系统提供了比早期的更加有用的模型，这种维度更占据优势"（2013:24）。整合传统表达行为的多重维度并非易事，而是姗姗来迟的工作。[①]

树上海盗的故事

我研究民间叙事的计算方法是始于十余年前的一场机缘。我拜访过一位非裔美国人奥斯卡·巴比诺（Oscar Babineaux），他住在南路易斯安那州一个小镇上，作为传统韵体诗的天才表演者，巴比诺给我表演了一段，美国民俗学者有时把这称为祝酒词。巴比诺能自如地在文本间转换，从诗歌转向笑话，之后转向传说和轶事，在多样化的文本连接之间，他给我讲了两个丢失的宝藏的传说。一个传说包括了讲述者和坐在树上海盗之间的视角交换，这个故事结尾不好，海盗诅咒了故事讲述人。我不知道这个奇特的故事是怎么形成的，巴比诺还热情洋溢地给我讲，这个故事追逐了我很多年，直到我决定在看起来相当现实的人的语料库中寻找到关于它的解释。

巴比诺住在莱恩镇，有 8500 人定居于此，小镇位于美国路易斯安那州拉斐特市西边十里地。[②]这个小镇位于路易斯安那大草原的中间，最显著的特征是南太平洋的铁路线横贯于此。（参见图 2）这个小镇很有意思，原因很多，其中之一是 2000 年人口普查时，小镇上有 34% 被确认为是非裔美国人，轻易地成为该区域非裔人口比例最高的美国小镇。小镇上黑人和白人的比例，远高于其他小镇常见的比例，即 1:5，有一种较有说服力的解释认为，这源于 19 世纪末期土地的可获取性。历史学家卡尔·布拉索（Carl Brasseaux）从拉斐特和教会角附近地区固定下来的人口数来推测，莱恩镇是外来非裔美国人移民的接收地。[③]比起周边其他区域的非裔，他们在莱恩镇有时会拥有一块土地中较好的份额，这种情况还较为常见。这个事实在稍后提到的巴比诺的故事中发挥着一点作用。

图 2 莱恩小镇位置

巴比诺和他的妻子住在一幢整洁简朴绿色房屋中，他们生养了两个女儿。他充满青春的活力，胡须整洁，体格结实，讲话精力充沛。他是个有自我认识的讲述人，巴比诺自己出生于虔诚天主教家庭并有此信仰，但他不久前刚邀请了一对摩门教传教夫妇来家中，"我告诉他们不要试图改变我的信仰，我也不准备改变他们的信仰，但他们可能会教给我一些我之前并不知道的事，也许我也可以教给他们一些他们所不知道的。"当我到达他家门前台阶时，他用同样的热情和好奇欢迎我并很快交谈起来。巴比诺其

① 唐克里尼是推动计算民俗学研究的工程师。他的工作主要是尝试为丹麦民俗建构基础结构，即可见的模型。但是遵循他的路径他介绍了大量的主题和方法，值得更多的讨论（参见 Tangherlini 2010，Abello, Broadwell，and Tangherlini 2012）。

② 我接近巴比诺是由于他的女儿为我播放了一段录音，她录下了父亲在一场仪式中间所表演非裔美国人的祝酒词。一些读者可能知道，很多当代非裔的民间文化是城市文化的附属品，但是他们还具有深厚的农业之根源，像当代农业实践一样，值得深入研究。

③ 个人邮件交流。

实是当地非裔美国人社区中因表演力而闻名的人，擅长表演多种多样的话语文类，当地一般称为"大杂烩"。人们会告诉你，奥斯卡·巴比诺真能扯，他们的意思是他讲得棒极了。①

对于巴比诺和这一社区的大多数非裔来说，大杂烩包括了多种形式，如祝酒词、笑话、夸耀、辱骂等，这是其中一些能被叫出名来的文类。在莱恩镇，最常有的是下午的社交聚会，通常在家里，男人们，有时也有女人和孩子，三五一堆聚在树下或冰箱周围，或者晚上在门廊前面，或有时在门廊里，为更突出的讲话准备好通常的社交氛围。奥斯卡·巴比诺从多种资源中为自己的素材库拾遗收集，周末他有时跟着父亲从一家到另一家，有时在树荫下和甥侄们消磨时间。

奥斯卡·巴比诺表演了大量的祝酒词并给我讲述了很多笑话、传说和轶事，我在其他文章中有所讨论。在我收集到完整的表演段落中我想提到的是，树上海盗的故事是巴比诺讲给我的几个纪念性传说之一，即从第一人称的视角，报告了发生在过去某个不定的时间段内的事件。这个故事，像别的传说一样，并不能限于某个时间顺序或历史阶段来考证，有趣的是，考虑到整个篇章，传说中的变动是有暗示的，在某种程度上，是通过变换为第三人称视角，来主导祝酒词和笑话的讲述。轶事等通常遵循较为清晰的时间框架，置入了个体的自传：例如，他讲述过一些当他还是个孩子时的故事。

所有他给我讲的故事都是生动的第一人称叙述，对话丰满，充满戏剧性。大多数能提炼出一般在路易斯安那民间文化中所具有的母题，但是绝无仅有的是，我这里要提到的这个能说话的海盗。这则传说讲到，巴比诺家族的一个成年人在他的老家停下来，发现他的家人又在挖钱。他和家人们一起祷告，然后他和侄子一起拿水给屋外的人喝。这样一来，他们遇到了一个坐在树上的海盗向他们要喝的。起先，他们答应了，随后他们又拒绝了，这个海盗威胁他们，这时铁铲也飞到了天上，自己插到了树里。

正像我所说的，我家族的人都觉得不可思议。他们想挖钱和东西，我祖父还留给我们一些钱，他们就是挖这些钱。那天我们就出去了，我在工作，所以我能看到，我们在村边，像我们的财产。我就看到很多人穿着白衣服，我就很好奇。我说，"哦，天哪，每个人都穿成白的到底要干什么啊？我得瞧瞧去。"我就出去了，他们就告诉我，"你现在在工作，快回屋去，你知道，回去工作去。"所以我又回去了，去工作。所以呢，他们都在房子里。我们都在祷告，每个人都跪着祷告。他们弄了个打洞机在后院，挖啊挖（笑）。你明白了？在找这个钱，我猜。我们都跪着，嘿，我们在祈祷。就像在夏天一个矿坑中，像在这里，没有一丝风。他们让风刮过了房子。风特别大，我的姑姑那样抓住了门，她的腿被刮起来。风有多大啊，就在房子里。

然后他们说，他们选了我，我的侄子，就是那个我告诉你啥都能扯的，和我的小侄女去拿些水给后院的工人，那个人还在工作。所以我们就走过去了，我们拿水顺着房子走过去。

然后我侄子说，"看啊，嘿，你看到树上那个人了吗？"

我说，"真见鬼，树上没有人啊"。

他说，"有呢，嘿，他就坐在那个树杈上"。

我说，"嘿，我没看到有人啊"。

现在我吓坏了。

天，我真没有看到人啊。

但是他看到了，你知道。

所以他说，我说，"那人啥样？"

"是个男的，"他说，"他穿着海盗的衣服，嘿。"

他说，"他戴着海盗帽呢，还穿着海盗的外衣。"他开始和他说话。

① 这里简要解释一下当地语言中的大杂烩，是称笑话，讲故事和吟唱诗，还有相互辱骂和互相逗嘴，像母女那样。

树上的人开始和他说话，正当他和我说这些时，树上的那个人告诉他：闭嘴别告诉我这些。
所以他和我说，“嘿，看他就在那，你看不见他吗？看他就在那个枝子上。”
他说，“他也想喝点。”
你知道，因为他们已经在做：他们在后院放了一个大碗，就在树下，里面装着一些酒。你明白吗？
我不知道是否是阳光能够溶解掉，它就不见了。
好吧，他说了他说的话，“嘿，他还想再来点。”
我就说，“该死的，别跟我说这些。”
我想回屋里去。
我说，“我看不见那上面有人。”
我们就继续走，我们走出那块，我们给他们拿了水，然后回去。
看着他。
他说，“看你，你个臭小子。”
他说，“你就不能再给我一碗酒，啊？”
他说，“你有点像我啊。”
他说，“你看到这个猪腿了吗？”
他说，“你有点像我啊。”
他说，“因为你们在这，你们要丢掉一些东西。”
嘿，这真叫人有点怕。我们开始快步走。
当我们快到屋子时，我跑起来。一把铁铲，嘿，从屋后过来，我是说，全力冲来，那把铁铲插在树上，深深的。我们不得不用斧子把它取出来。它卡住了，你知道，铁铲，很难卡在什么地方，那把铁铲插进了半个树干那么深。

做这个记录，我努力将讲述者原本的话语尽可能的记录在手稿中。一些读者可能会注意到，很多处用了“他说”引起一个段落。正如我所指出的，我想无论是在口头讲述还是简单的口语化维系的连接链中，如果丢失这种字句开头都是错误的。而且，我还想指出，特别是在“他说”这种用法里，不仅是此处讲述还有其他表演中，都有传统化的维度在发生作用，正如交流过程，通过直接或间接的引用，使得多样化权力得到某种分散。（这种在口头话语中的分散所起到的作用也会显现在对民间叙事的计算方法等更大的讨论中，有机会再予以讨论。）

计算一则传说

所以，那个海盗在树上干什么，为什么他要威胁这群非裔美国人？答案更加有趣，它带领我们进入相关文化和历史当中去追寻。要回答这个问题，要尽力探索南路易斯安那围绕宝藏故事的观念的大文化网络，我收集了两类文本：民俗学者收集的口头传说，以及公开发表在互联网站上的关于宝藏猎奇的传说。一共有 36 个故事：20 个来自当代民俗故事集，16 个来自各种网络资源。由于两类文本似乎具有多种可归类的民俗维度，也许最好的方法是将他们分成：一类文本来自民俗学的口头故事集，另一类文本来自网站的网络资源。这种指定仅仅是工作上的区分而没有任何超越当前讨论的意味。

口头故事集，现在起我就这么叫，由 20 个传说组成，都是像我这样的民俗学者收集的。4 个来自于巴里·让·阿西来的《路易斯安那州法裔和克里奥尔人的民间故事》；4 个来自我大学的本科生，杰弗里·布鲁萨尔；2 个来自我在莱恩小镇对非裔美国人的田野调查；9 个来自“交换故事”文集项目，另有 1 个十行文本不在此文集中，他的作者慷慨地给我使用。在建立网络故事集的过程中，我优先选用能够提供

路易斯安那州宝藏并似乎比较活跃的虚拟的账户，同样，我们似乎还可以称之为“社区”。[①]“活跃”在这里意味着对任一给出的线索有较多的一系列响应，那些账号他们自身也主要是进行一些讨论，这样我能尽最大可能来计算这些故事，在某种形式上这是更大群体的认可。[②]

逐字谈到文本，这两个集子算势均力敌，16个口头文本的表现，就我所知，相当严格地转写了原本口头的话语，这4个布鲁萨尔的文本与原本的讲述相当贴近，但是他们最可能被视为比表演还有过之的文本。我想，这种差异对于那些文本所展现的情节并不是最重要的，但考虑到整体篇章时确实有点影响。例如，讲述者称布鲁萨尔曲解了口头故事集的语词差异性，将它视为与网络文本相近的内容。（参见表1和2）

表1：口头传说的长度和差异性

文本	总计	独特性	差异性
LOH 164	1024	339	0.33
LOH 165	904	308	0.34
LOH 160	760	306	0.40
LAU 14	676	218	0.32
LAU 13	384	185	0.48
LOH 157	367	180	0.49
ANC 88	333	148	0.44
LOH 162	331	154	0.47
LOH 161	295	141	0.48
LOH 159	279	149	0.53
LOH 163	209	120	0.57
LOH 162b	194	114	0.59
LOH 158	193	111	0.58
ANC 91	174	111	0.64
ANC 89	153	86	0.56
ANC 90	138	75	0.54
BRO 3	136	90	0.66
BRO 2	122	79	0.65
BRO 1	117	74	0.63
BRO 4	67	50	0.75
平均值	343	152	0.52
最小值	67	50	0.32
最大值	1024	339	0.75

① 我还没看到批评的讨论社区和社团之间差异，韦伯指出他们是两种完全不同的实体，在线社区的属性有一些更像是社团。

② 这个集子有三个来源：11个文本是今年4月13日复制自宝藏网站点，自称是“宝藏猎奇原生故事站”，同一内容有三个名字指向，即“失去的宝藏在线”，另见安东尼·贝里（Anthony Belli）出版的《失去的宝藏》2010年11月卷；2个文本来自于“宝藏发现之梦”，2012年5月8日，版权所有人为吉姆·罗沙（Jim Rocha）；在这16个故事中，来自宝藏网站点的故事更具民俗属性而打动了我，但是这种差异最好留待日后更细密地探索和分析。

表 2：网络文本的长度和差异性

项目	总计	独特性	差异性
TN 8	3081	1144	0.37
TN 5	2749	1089	0.40
TN 3	1957	560	0.29
TN 2	1112	477	0.43
TN 9	1050	439	0.42
TN 7	976	520	0.53
TN 4	913	437	0.48
BELLI 1	850	397	0.47
TN 10	769	390	0.51
ROCHA 2	482	284	0.59
ROCHA 1	447	252	0.56
TN 6	425	211	0.50
BELLI 2	303	183	0.60
TN 1	300	177	0.59
BELLI 3	253	145	0.57
TN 11	155	109	0.70
平均值	989	426	0.50
最小值	155	109	0.29
最大值	3081	1144	0.70

粗略看下表格显示出语词的差异性，正如文本中语词的整体数量所测算关于这些语词的每一次重复使用的比例，在口头故事集中，比例较低于网络故事集，但是两者都超过了大量口头文类的平均值[①]。如果在这组基本的统计数字中有什么值得说明的，那就是语词的差异性在口头故事集中比在网络文本集中更受限定。最突出的差异是，仅就表面来看，网络收集到最长的文本是口头文本的三倍，但是在眼下如此小的样本库和语料库的初期建设中，它应该被记住，这并非仅为促进讨论而提及的轶闻信息。要得出任何结论都还没有足够基础：我在表 3 中放入了一些我这些年收集的来自口述史的基本统计数字，仅对可能由此属性而影响到的长度多样性和语词差异性提供参考。

表 3: 作者收集的口述史样本统计值

平均值	111	69	0.64
最小值	43	34	0.55
最大值	196	108	0.79

暂时抛开网络文集，因为我想集中于实际的口头传统资料和变动性，甚至是对口头文集最简要的检验都揭示出文本在长度上的巨大变化：最小的文本 -BRO4，在口头文集中只有 67 个语词。最长的文

① “语词差异性是测量文本中使用了多少个不同的词，语词密度是测量文本中语词类型的比例，如名词、动词、形容词和一些副词等。这两种测量都有利于更容易的操作，也更实际地应用于计算机对大数据集的分析，而且，语词差异性和语词密度都显示出在书写中比在口头中有显著的增高”（Johansson：61）。参见 Ure 1971，Halliday 1985。

本-LOH164，有1025个语词。然而，对最小文本的细致检验，揭示出它多半是对一则传说表演的有关报告：

听到这
我不由得想
海盗在南路易斯安那埋宝藏
我学了这段历史
海盗在地上埋宝藏
要牛仔帮他看守
守护者被抢杀
他的灵魂便永远守护
也许这两个故事有关连

请注意，布鲁萨尔的第二个文本BRO2，相当接近于文集中其他对此传说的全篇描述。我这里展开一下：

这亲戚的丈夫
他走去那片木林
他的兄弟也跟着，在找柏树苔藓
为靠枕和篮子寻找软垫
这天在回家的路上
他们发现树下有块石板
这很古怪
因为他俩撬不动这块石板
它似乎嵌在地上
他们觉得奇怪
这块石头有点不对劲
两个人吓坏了
赶紧爬上树
再向下看，石板不见了
他们从树上跳下来
跑回家

BRO2有122个语词，还是相当短小，但是还短不过来自阿西来收集的一个文本：也许这些文本的长度是具有代表性的讲述这一文化中讲述宝藏故事的最低临界值。这个文本是我们讨论的焦点，树上海盗的故事，有652个语词，是口头故事集中较长的文本，讲起来要用3分钟。

暂时抛开计数这件事，我们之后会回到这篇故事，让我们先对这些文本进行定性辨别。也许能将口头文本从网络文本中辨别出来最关键的特点是，后者倾向于关于宝藏的来源，而很多口头故事，就像我们传说所讲的，关注于寻找宝藏的经验，对于来源一带而过。两个文本都有这些内容，在这个项目中我将此视为证据，我们得到了形态学上的经验，这里称为“τ成分”（τ），在口头传统中发挥作用；来源称为“α成分”（α），在书写传统中发挥作用。每种成分都可再细分成更小的部分，我不太想称为功能，不过在叙述体系的描述过程中可以视为有与此相同的价值。

发现我这里收集到的故事具有同样的形式，我承认有点震惊，如上所述，我称之为“τ-α”。在这种情况下，(τ)是寻找宝藏的经验，常出现在个人的过去史，(α)是宝藏的模糊或丢失，更多是非个人的过去史，或者多少有点传奇式。不是所有的文本都有这两种类型的内容，但我有把握这是类似的，单独的传说也是具有经验或来源两种成分，在某种程度上另一部分内容的存在是可以分析性地推论得出的。这是重要的差异。我想都市传说的相传不会包含全部组成部分，但是让我觉得更有趣的发现是我有 20 个文本，在这 20 个中间，有 14 个是由寻找宝藏组成的(τ)，4 个仅有宝藏是怎么来的(α)，2 个是二者兼有的。[①]（参见表 4）

表 4: 口头故事集中的 τ-α 成分

成分	发生率
τ	14
α	4
τ-α	1
α-τ	1

在口头故事集中，仅由 α 成分组成的 4 个文本，包括了 1 个布鲁萨尔的收集文本（BRO4），3 个来自交换故事项目（LOH157，LOH162B，LOH163）。2 个兼有二者的文本却有完全相反的组成顺序，一个是时间顺序的 α-τ（LOH162），另一个是经验顺序的 τ-α (LOH-160)。

这里每个成分的例子都是有用的。我们先从主导的叙述成分开始，来自阿西来的《法裔和克里奥尔人的民间故事》：

我去和马雷罗的一个老人碰面
他给我讲了一个故事
他和一些人找宝藏的故事
有一个掌控者
带着圣经掌控幽灵
他们到了地点
他们看到一匹大马穿过树林，上面骑着一个人
当他下马时
发现那匹马背上没有人
是一条狗
他说狗过来了在他腿上来回蹭
他说它咆哮着
他说他知道狗在碰他
但是他感觉不到
就好像只是一阵风

① 意识到这点，我有很多个这样的故事，而且大多一样，我又重复地检验了这个过程：我是不是还有遗漏的？是不是我的直觉先于我意识到一种模型决定我在故事集中所选择的文本？我可以说我并没有这样做。我继续寻找所有以某种方式提到宝藏的文本，无论是说钱还是黄金还是硬币还是其他的宝藏。我逐渐拓展了我的研究范围至可能有某种未知的收获或赏金的故事。但是这种拓展只会增加更多样的文本结构，而非缩小。

而且他说他们都开始拼命跑
他丢了帽子和眼镜
还有他撕开自己的外衣
而且甚至这个领队者也跑
那之后他再也没有看到他的圣经（ANC89）

首先要注意的是这个文本的开头——“我去和马雷罗的一个老人碰面，他给我讲了一个故事。”20个文本中有14个用这样的开头，具有代表性地使得讲述者成为故事情节链上的一环，这种文本的特色，文本制造中的特色应获得更多的关注。在某些个案中，开头需要建立起历史精确性，或者，是真实的混合型，试图协调在作为故事接收者和给予故事内容历史真实性两者之间的一般理解上存在着的区隔。

开头之后，这个短文本就开始进入τ成分，它提供了τ成分两种不同类型中的一种类型，即实施者去挖宝藏。另一种类型在来自“交换故事”的下文中可以看到：

在拉斐特市附近的杜森小镇，
90号公路上，
当你从这边来快接近镇子时，
你能看到有一个小天主教堂在马路左边，
有一个人叫朱迪斯，他很热心公共事件
他让人耕地
犁碰到了东西
他们打开它
是一大罐珠宝和金币
都是很早以前的
是法国金币和早期美国金币
他们把它又埋起来
他们说属于拉斐特
他们不知道他是谁
但是这个黑人用一部分钱建起了这座天主教堂和学校
七年之后，我猜，妒忌，让它烧毁了。
他又重建了。
七年之后，它又烧毁了。
他再次重建了。
他还剩多少钱我不知道
这是个例子
明显有事实根据的。[①]

在第二类型的τ成分中（τ2），实施者没有挖宝藏，尽管他们常有其他形式的挖掘行为，通常使得他们实际上找到了宝藏。这种事件表明τ成分有如下两种形式：

τ1中实施者来到一个地点，挖宝藏，并遇到幽灵。

① 为了行文方便我把文本的开头去掉了。

τ 2 中实施者来到一个地点，执行农业工作（犁田、牵牛、割苔藓、打猎）并找到了宝藏。

有意思的是，在两种变形中实施者前往的地点都常常是一致的：树林，15 个文本中有 6 个或直接指出或暗示。（参见图 3）

如果我们比较两种 τ 成分，将其视为一系列的叙述状态（或功能），那么我们就能看到，当 A 完全一致时，B 通常表现得相当类似，C 会颠倒。对此差异唯一能解释的是人类的意向性。如果你去一个有意要找到宝藏的地点，你不是仅为了找到它，你还可能发现自己撞到幽灵不得不逃命。然而，如果你到一个地点有意去做某项工作，这种工作在有些文本中时常与日常财富的增加相连，像是收集松塔，然后顺利找到了宝藏。如果你没有找到宝藏，那么你会遇到坟墓或是石板，在其他文本中这些都和宝藏相关。

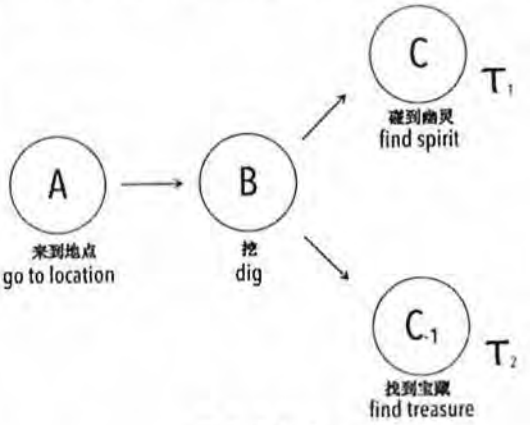

图 3 成分图

τ_1: A ⟶ B ⟶ C

τ_2: A ⟶ B ⟶ C_{-1}

小图示说明：这些简图中的箭头只是表明连接的顺序，可能出现在叙述的实例当中，并不是任何形式的因果关系。（实际上已经融合起来了）

两种形式的 τ 成分都有附件的复杂性，有很多文本显示出对事件有趣的重复，可设想为一个圈，通过多样的表述，像芬雷森在他考察普洛普形态学所应用到的多样叙述集。

口头故事集中的四个文本，只在布鲁萨尔的 BRO4 中有 α 成分叙述，2 个来自“交换故事”项目的文本（LOH157，LOH163），菲舍家中黄金的传说是个很好的例子：

我要给你讲的故事里这个人
我知道他实际存在过
因为我还在他坟墓前玩过
他被埋葬了，还埋着呢
就在我们住的地方
他被埋在院子里
就是我住过的
他们用柏树桩围起篱笆
那时我已经长大知道事情了
篱笆断开了
但每根桩子还完整
有个人据说叫菲舍
显然不是法裔名字
可能菲舍和他妻子以及妻子的儿子，叫比利的，住进那幢屋子
他们从哪里来
没有人知道
这个故事是
这是传闻和猜测
他是个银行劫匪

他搬进那幢屋子就消失了
他酗酒
每次他去镇上
他都会醉
这本来是教堂角，最近的小镇
他骑上马
去了小镇
回来就醉
还打比利
一天下午他喝醉回来
比利对他开枪
杀死了他
他的妻子和比利把他埋在那里
晚上天黑了
他们坐马车走了，可能还带着很多金子
他们来到让小詹尼斯的家
这不是狼人
这是他的儿子，还活着
房子还在那里，屋子不在了地方还在
当他们到达哪里
她把小让妮看成可依赖的人
她就在黑暗中停下来

这总是在暗夜里！
他告诉她
"如果你想在夜晚穿过树林
你就会被抢劫
为什么你今天不留下来
明天就可以走。"

可能她回到了密西西比
那个晚上，可能，她把钱埋在小让妮的家，很多黄金
多得惊人的黄金
她再也没有回来
所以黄金还在那里
我一个朋友告诉我
那是真的
因为所有的醉汉都有很多钱可以埋！
那是我对那故事的感觉

在表面上，它更像一个闲谈故事除了偶尔提及黄金埋藏在某处。但是，还有个死去的人和宝藏有关，传说中的转移，是α成分的一般特点，故事讲到了宝藏的来源。

关于这些成分还可以有更多的讲述，在故事集中他们如何贯穿，但是今天我们关注的是海盗如何坐在树上，还有他为什么威胁非裔美国人。想到这点，当我们关注来源故事，我们发现如下情形：

在第一个文本（BRO4），海盗枪杀了牛仔，使得他的灵魂能够保护宝藏。

在第二个文本（LOH162），家族的财富先是被用一桶面粉埋好，后来这些钱被家族成员和两个斯拉夫人向西转移，家族成员死了被斯拉夫人用钱埋了。

在第三个文本（LOH164），一个奴隶答应了要去寻找家族财富就被杀了，这样他的灵魂“能够继续守护金钱”。

在第四个文本里，故事散漫，除了有提到金子被埋在某处（LOH157），一个死去的人与财富有关并有转移，是 α 成分的一般特点。

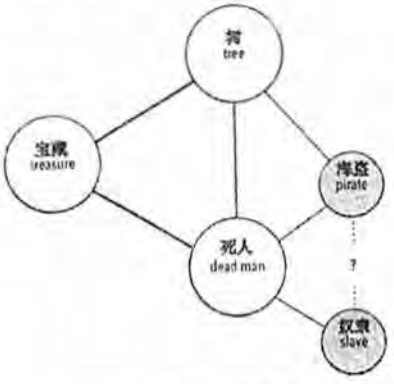

图 4 关于宝藏来源所设想的网状草图

设想的网状草图会包括以下关联：财富、树、死人、海盗、奴隶，那么文本是否能够让我们建立起从奴隶到海盗之间的连接，而不是说这两者在语料库中发挥同样的作用？（参见图 4）

另外一个文本指明了路径：LOH163 讲述了著名的海盗让·拉斐特跑去德克萨斯州来逃脱因自己做错事而被抓捕，也提到他做对的事。拉斐特住在了加尔维斯顿（德克萨斯州东南部），但在路易斯安那州仍然有生意，一天他不得不放弃萨宾湖上的一条船，它满载着财宝。在湖底沉船上能够发现哪种财宝呢？历史记载证明了海盗卷入了奴隶交易，其中一种记载发布在网络上，收集到语料库中：

到 1817 年，让·拉斐特及其前任路易斯·奥瑞的私掠船从古巴沿岸俘获了大量的西班牙奴隶人。海盗在加尔维斯顿岛的奴隶收容所、围栏等常常都超容，挤满上千非洲奴隶。很多买主只花一英镑就买走奴隶，鲍维三兄弟：约翰、雷津和詹姆斯，是海盗最好的买主。1853 年，约翰·鲍维在“德宝杂志”上记载他的兄弟，通过黑河口到萨宾湖或者走卡尔克苏到查理湖上铺开非法奴隶交易，在路易斯安那卖掉 1500 个非洲人而在两年内实现获利 65000 英镑。[TN 8]

这里提到的记载揭示了奴隶和财富之间经由海盗的明显关联。奴隶就是财富，他们在新大陆的辽阔区域内被交易。奥斯卡·巴比诺拥有这个有力的关于历史教训的传说，他没有提及这些，也许也不知道直接的关联只是感到有种连接。

为什么计算语词

为了说明和探索，旧的民俗学观点与方法论像形态学和主题关联网似乎也发挥点功用。我这里使用一个小文本集作为分析的背景。我围绕一个特定的文本收集到的这些内容是为了推动处理进程。沿着这一路径，我构建了一系列表格和图示，这样做的目的是要了解，基于人类定性研究的理念如何能继续扩大此项研究。很多学科已经产生关注于探索定量研究的领域，这很可能是由当前时代充足的计算能力决定的。

人们仍然会想，“为什么计算语词？”也许有人会回头去看这篇论文迅速给出的一系列数字，这些数字只不过是传说文本集的初步分析，只是探索叙事计算方法的第一步。要说明这些就是一张表格，但它只是 16 个传说文本的简单长条图。请记得每一个文本，都是由确定的民俗学者收集的（所以我觉得原本的口头文本也比较可靠）。最长的文本超过 1000 个语词，最短的，大概 100 左右。增加十倍只是一个很小的量级，但是对于进一步研究已经提供了足够的宽度。

这个问题的初步回答是简单的：我计算语词是因为，我想知道是否可能用100个左右的语词创建一个故事世界。如果可能，我想了解这怎样发生的。考虑到大量的文献形式的长度，1000个语词已经是相当简洁了，那么100个呢？每个词都装载着不可思议的巨大力量：有时变得更加精彩，当人们意识到只有一半的语词在用法上是具有独特性的，甚至在这么短小的文本中！如果我们转回到ANC89（上述的），我们数出153个语词，但是，这些词，一个单独的词，“他”，用了12次。接下来用得最多的9个词也都很好理解：和，一个，是，这，它，他的，说，去，他们。所以一篇文本用得最多的10个词并不能说明故事本身，除此之外，也许还有单数“他”，是与群体的“他们”相抗衡的（只有当我们进入接下来10个常用词，那些文本中出现了2-3次的词，我们开始能有所感觉，故事可能说了什么：人，狗，和，当，走，那里，看到，离开，马，掌控）。

这怎么可能？来自已经这么短小的文本中这么小的子集语词如何使故事进展？我想，这是真正的问题。语词计数只是这个方向的第一步，却是重要的一步，是我们作为民俗学者没有做的。回想一下那些20世纪重大集成项目中编目的所有文本。而且，我们收集的所有文本还处于言语民族志的保护下。这是令人印象深刻的巨量工作。我们进行了一些综合识别，总的来说，大多是关注差异的。所有那些差异，当然，相当引人注目。但是关注差异，我们也就错过了全面谈论人类本性和文化的机会。

语词计数的推动力，对我来说，仅是迈开了一步，来充分理解人们如何借助世界思考，如何借助外物或他们自己制造的事物来思考。在文本的例子中，他们将一个词跟着一个词逐一排成一串，通常是置于更大的话语流中，这可能有利于或不利于文本的制造。除了这些复杂性外，人们处于不同言语行为语境时会决定新讲一个文本，将一个词放在另一个词后，按照他们的预期和操控形成序列，直到它们都符合某种类型和结果，就像话语领域的阿特洛波斯（希腊命运三女神之一），结束词串的生命。

那么，语词计数，仅是迈开了一步，为更充分地理解语词的数目，乃至具体词汇和词的顺序。为什么是这些词而不是另一些？用在此处营造了一个故事世界的这些词，之间的关系是什么，故事世界的情节和他们所设置的人类世界之间的关系又是什么？简言之，150个语词告诉我们的词语、思想和行为之间的关系是什么？

之前时代的民俗学成果中最大的目录往这个方向走了一步，尝试勾勒框架，大多是书目文献而且是间接的目录式的，起初多种文本被收集在文献学项目里，同时，斯蒂斯·汤普森让伟大的旋转木马转起来，他是用一张张三乘五寸的卡片来编辑母题索引，然而一些学者和科学家开始用计算机处理，来汇编文本的统计数字，那时他们这样做又慢又昂贵。①

对于很多人文主义者来说，统计数字或是谜团或是仇敌。对很多人而言（有充分的理由），它再现了一种数学运算的制度，它自身有点神秘，常常被用于总结一种情形或一群人，有时还需要更精细的分析形式。在这篇文章中，我不会为这种语境中的使用来辩护，我对辩护或讨论也没有兴趣，统计式的更大转向是如此多知识产物已经采用的形式。我这里仅对我自己的童年做一点修订，也许你也会：并不是其他人做了什么我们就也要去做什么。

我很能理解人文主义者试图站在多样的立场并要喊出“我们会被压碎成数字终结至此”。在此我建议，在我们前方隐喻的路上，压碎会一直进行着，有没有我们的努力它都照样进行，不是要将压碎人性化而是要让它充满人性，以致于它能够很好地变成一门新的科学，某种新的学术成果，不仅使其他人感兴趣，也使我们自己能够感兴趣。

统计数字的核心需求是你必须把这样的信息——也许是一个关于宝藏埋在哪里的简单小故事，也许是十来个这样的故事，或者是几千个故事——转换成数据。但是这种转换量只是简单给问题核心的客体

① 斯蒂斯·汤普森坐在直径40英寸的旋转木马屋里的图景完全归功于亨利·格拉西。开启旋转木马需要进入一个有控制箱的房间，打开控制箱需要有一个选取码，用选取码来比喻汤普森的编目工作。

设定值，通常是数字，但是它们不一定需要。分析者定义了问题，分析者设定了值。民俗研究已经在故事类型的数量、母题的数量、甚至在我们描述一个特定文本的语境化过程时，这样做了。

那么为什么计算语词？这样做，明显的理由是，简单探索文本和文本性，去满足我们对于人类表达的基本层面的好奇心：文本中的语词数量，文本中出现的语词丛簇（或配置）还有那些总是在特定形式的文本相互连接中出现的语词（共现）。第二个理由是，推进这种方法将有助于我们发现文本之间新的关系，这些关系用更为传统的研究方法是难以发现的。发现，实际上是思想对自身的索引，是自然语言处理为何如此用力的背后主要原因，这正是本文开头所指出的。最后一个理由是，通过对民俗文本新的观照并探寻文本间我们以前没有注意到的关系，将会引领我们走向新的知识形式，这些形式无需取代当前获取知识的路径，而是要重新定义并拓展这种途径。①

引用及参考文献：

Abello, James, Peter Broadwell, and Timothy R Tangherlini. 2012. "Computational Folkloristics." Communications of the ACM 55 (7): 60. doi:10.1145/2209249.2209267.

Ancelet, Barry Jean. 1994. *Cajun and Creole Folktales: The French Oral Tradition in South Louisiana*. Garland Publishing.

Barthes, Roland. 1981. *Camera Lucida: Reflections on Photography*. Tr. Richard Howard. Farrar, Strauss, Giroux.

Carron, Pàdraig Mac, and Ralph Kenna. 2012. "Universal Properties of Mythological Networks." EPL (*Europhysics Letters*) 99 (2): 28002. doi:10.1209/0295-5075/99/28002.

Laudun, John. 2012. "' Talking Shit' in Rayne: How Aesthetic Features Reveal Ethical Structures." *Journal of American Folklore* 125 (497): 304–326.

Lindahl, Carl, Maida Owens, and C. RenŽe Harvison. 1997. *Swapping Stories: Folktales from Louisiana*. University Press of Mississippi.

Saxon, Lyle, Edward Dryer, and Robert Tallant. 1945/1987. *Gumbo Ya-Ya: A Collection of Louisiana Folk Tales*. Reprinted by Pelican Publishing Company.

Tangherlini, Timothy R. 2010. "Legendary Performances: Folklore, Repertoire and Mapping." *Ethnologia Europaea 40* (2). Museum Tusculanum Press: 103–15.

Tangherlini, Timothy R. 2013. "The Folklore Macroscope." *Western Folklore* 72 (1): 7–27.

Word, Christine. 1988. *Ghosts along the Bayou: Tales of Hauntings in Southwestern Louisiana*. The Acadiana Press.

Victoria Johansson. 2008. Lexical diversity and lexical density in speech and writing: a developmental perspective. Lund University, Dept. of Linguistics and Phonetics *Working Papers* 53: 61-79.

[责任编辑：张礼敏]

① 对有兴趣的读者而言，源代码知识库（https://github.com/johnlaudun/CountingTales）里有我工作过的材料，有文本也有Python 语言脚本，还有一些其他材料和研究。由于版权限制，我这里无法呈现阿西来和林达等编辑收集的所有传说文本，但我想将来能和他们协商好使用。