文章编号：1007-1482（2018）01-0117-0124
DOI:10.13505/j.1007-1482.2018.23.01.016

·体视学方法应用·

汉英引号用途对比和汉字构词歧义:借用抽样方法的试点研究

杨正伟

（川北医学院 形态定量研究室， 南充 637000）

摘 要：借用体视学常用的等距抽样方法,本文试着对比了汉英书面新闻语料中引号的用途,分析了汉语书面新闻语料中汉字构词歧义的出现情况。结果表明,用作直接引用话语的引号,在汉语和英语中的使用频率分别为28%和88%;4.3%的汉字出现构词歧义,即这些字与其前面或后面的字形成了非原文含义的词语。此外,对未来这样的定量语言学研究,本文讨论了其抽样相关问题。

关键词：汉语；英语；引号；构词歧义；抽样估计；语言学；定量研究
中图分类号：H136，H146，H315，N3 **文献标识码**：A

Usage comparison of Chinese and English quotation marks and word-forming ambiguity of Chinese characters: A pilot study with sampling estimation

YANG Zhengwei

（Morphometric Research Laboratory, North Sichuan Medical College, Nanchong 637000, China）

Abstract: By means of systematic sampling commonly used in stereology, the present paper tentatively compared the usage of quotation marks in written Chinese and English news materials and analyzed the word-forming ambiguity of Chinese characters in written Chinese news materials. The results showed that 28% and 88% of the quotation marks were used for direct quotation of discourse in the Chinese and English news, respectively; ambiguity occurred with 4.3% of Chinese characters, which formed a word or phrase, with the characters ahead or behind, that had a meaning different from the original meaning in the text. In addition, sampling related issues with such quantitative linguistic studies in the future were addressed in the paper.

Key words: Chinese; English; quotation marks; word-forming ambiguity; sampling estimation; linguistics; quantitative study

收稿日期：2017-12-13
作者简介：杨正伟（1963-），男（汉），教授。研究方向：形态定量研究。E-mail:zwyang@nsmc.edu.cn

0 序言

生物、材料等组织结构(细胞、晶粒等)的形态定量特征(体积、表面积、直径、数量等),一般用体视学方法进行抽样估计;所涉及的抽样方法,常用的是等距随机抽样[1]。应用这样的抽样估计方法,笔者最近试着定量研究了语言结构(字、词等)的某些形态特征:比较了中英对照书籍里的字、词数[2],确定了英汉资料里汉语字汇与英语词汇的数量及其与字符总数的回归关系[3]。笔者以为,这样的方法还可用于语言结构的功能(用法)特征研究。就是说,收集一定语料(总体),从中随机抽选出字、词、句或标点等(样本),然后一一分析所选结构的功能或用法,以此反映整个语料的情况(从样本估计总体)。而且笔者以为,抽样估计在语言用法研究中可能具有更大的实用价值。为此,本文试着抽样估计了语言结构的某些功能特征:汉英书面新闻中引号的用途对比,汉语书面新闻中汉字构词歧义的出现情况。希望这样的试点研究,能给未来相关研究提供内容与方法参考。

1 试点研究一:汉英引号用途对比

1.1 语料与方法

于 2005 年 2 至 3 月期间,登陆中国网(http://www.china.com.cn/chinese/),下载首页上显示的新闻要闻(汉语)共 123 条,只要题目和正文,拷贝至 Word 软件,排成刚好 100 页。于 2003 年 4 至 10 月期间,登陆雅虎网(http://www.yahoo.com/),下载首页上显示的由美联社(Associated Press)发布的新闻(英语)共 207 条,只要题目和正文,拷贝至 Word 软件,排成刚好 100 页。

分别在如此获得的汉语和英语新闻语料的 Word 文档中,抽选每页中出现的第一个前引号,然后分析该引号的用途,即分析该前引号及其对应的后引号在原文中发挥的作用,确定其用法分类(下面)。总共抽选了 100 个汉语引号,但 101 个英语引号,因为有 1 页中首先出现的前引号是连在一起的双引号和单引号的前引号(表 1)。

根据我国从 1996 年开始实施的国家标准[4],汉语引号的用途有三类,分别用以表示“直接引用”、“特殊含义”、“着重论述”的话语。英语引号的用法[5],与汉语引号的用法相似。因此,本文将汉英引号的用途分为三类,并试着根据引用话语的构成等特征进行了亚类划分,见表 1。

1.2 结果

利用 Word 软件的查找、替换及统计等功能测得,汉语语料的总“字数”(包括汉字、数字、英语字母等特殊符号,不包括标点符号)为 13.7 万,总共有 1930 个双引号,40 个单引号;英语语料的总“字数”(包括英语单词、数字、特殊符号,不包括标点符号)为 13.6 万,总共有 3329 个双引号,86 个单引号,另有 2219 个并非用作引号而是用作表示助动词省略、名词所有格、特殊复数形式等的单个单引号样符号(表 1)。

汉语新闻语料中,用作直接引用话语、表示特殊含义、表示强调陈述话语的引号使用频率分别为 28%、53%、19%,而英语新闻语料中这些用途的引号使用频率分别为 88%、10%、2%(见表 2)。汉语和英语相比,不同用途(三类)引号使用频率之间有高度显著性差异:统计学检验(χ^2检验)的 P 值小于 0.001。

1.3 讨论

笔者对英汉和汉语对照语料的试点研究显示,1 个英语词可能传递 1.04 ~ 1.72 个汉语字的信息[2]。因此,就所传达的新闻信息量而言,本文研究的英语语料可能大致为汉语语料的 1 ~ 1.7 倍。鉴于本文英语与汉语语料里引号总数之比为 1.7,英语引号的总的使用频率大约为汉语引号的 1 ~ 1.7 倍,即英语新闻中引号的运用不比汉语新闻的少。

该研究显示,英语新闻中,用作引用他人话语(A 类用途)的引号使用频率高达 88%,是汉语新闻中同类引号使用频率的 3 倍。因此,与中国新闻作者相比,美国新闻作者更倾向于用直接引用他人话语的方式来发表新闻。这种表达习惯上的差异也许说明,美

表 1　引号用途分类及例句

编号	用途	汉语例句	英语例句
A	**表直接引用的话语**		
A-1-1	引用多个句子	“这完全不是事实。讽刺的是……写在契约里。”xxx 表示。	“The outcome is not in doubt. That's … that,” he said.[b]
A-1-2	引用一个整句	“这是个好兆头!”xxx 说。	“What better … can you have,” he said, “than … to do.”
A-1-3	引用一个句子的片段(有 3 个以上的词)	xxx:xxx 是“支持……的首要国家”[a]	xxx said the lawmaker's comments “were specific to the Supreme Court case.”[b]
A-1-4	引用一个句子的片段(仅有 1 个或 2 个词)	xxx 在同 xxx 通电话时,xxx 要 xxx 停止“恐怖”活动……	xxx said any talk of an imminent threat was “pure conjecture.”
A-2	引用成语、谚语、常用诗词或表达等	“嘤其鸣矣,求其友声。”希望大家集思广益、畅所欲言。	… the national motto, “In God We Trust.”
B	**表特殊含义的话语**		
B-1	专有名词	……居然用的是“少林”商标。	“‘Commando’ is one of my favorite movies,” xxx said.[c]
B-2	特指对象、事件、活动、缩写、称呼、口号、术语等	……以邓小平理论和“三个代表”重要思想为指导……	xxx hoped xxx would implement the “road map” peace plan.
B-3	有特殊的含义	……企业很快“炮制”了一个广告额为 1.5 万元的假合同……	Saddam's “Shiite Thug” Is Captured[b]
C	**表着重论述的话语**		
C-1	表强调	……一些消费合同中并没有“消费者资料应该保密”的单独条款……	xxx smashed treasures that chronicled this region's role as the “cradle of civilization.”[b]
C-2	表词语或符号本身	……鸡由于音与“吉”相近,被人们赋予了吉祥的意义。	“The word ‘Polaroid’ is synonymous with instant photography,” said xxx.[d]

注:例句选自从新闻语料随机抽选的前引号所属句子。“xxx”表示某某人物或组织机构。[a]这个例句取自新闻题目。[b]这四个例句中单独的单引号样符号(例如 B-3 英语例句中的第一个“单引号”)不是真正的引号,而是用作表示助动词省略和名词所有格的符号。[c]该例句中被随机抽选的引号是连在一起的双引号和单引号的前引号,其双引号的用途属于 A-1-2 类,单引号的用途属于 B-1 类。[d]该例句中被随机抽选的是单引号。

表 2　随机抽选的汉语(100 个)和英语(101 个)引号的用途分类及频数

引号用途分类编号	汉语	英语
A(总数)	28	89
A-1-1	3	4
A-1-2	9	54
A-1-3	6	18
A-1-4	6	12
A-2	4	1
B(总数)	53	10
B-1	9	2
B-2	34	7
B-3	10	1
C(总数)	19	2
C-1	16	1
C-2	3	1

注:引号用途分类编号的含义,见表 1。

国记者可能更加注重新闻的真实性、原文感染力。

该研究也显示,汉语新闻中,一半以上(53%)的引号用于表示特指(B 类用途),是英语新闻中同类用途引号使用频率的 5 倍以上。这种明显的差异应该与语言本身的差异有关。汉语的词由一个或多个汉字构成,不分词连写,容易产生构词歧义(见本文的试点研究二),用引号隔开词语可以避免一些歧义。英语由词构成,且词间有空格,不易引起歧义。此外,英语还可用单词首写字母大写的方式等避免歧义。例如,表 1 中 B-1、B-2 用途的汉语例句中的“少林”、“三个代表”,如果译成英文 Shaolin、Three Represents,就不需要用引号。根据笔者的进一步分析,本文所抽选的 100 组汉语引号中,17 组引号内的汉语话语都可用首写字母大写的方式翻译成英语,即如果译成英语这 17 组引号都不必使用。

引号用途的分类,有时实际上并不容易。例如,

当作者仅“引用”一二个词时,有时就难以准确判断作者的本意是仅表引用(A-1-4 类用途)还是想表强调(C-1 用途),也许作者对此不在乎,也许作者本来就想如此“模糊”表达。又如,表 1 中 C-1 类用途的汉语例句中的“消费者资料应该保密”,看似引用的话语(A-1-3 用途)或表示特指(B-2 用途),但从其新闻原文看不出引用来源,也看不出特指什么。如果作者不是想强调“消费者资料应该保密”是将出台的个人信息保护法中应该包含的重要内容,用该引号就显得累赘了。不过,引用整句话的引号用途(A-1-1 和 A-1-2 亚类合在一起)的划分没有问题,表特指和强调用途(B 和 C 类合在一起)的划分也不会有多大问题。因此,即使若干亚类用途的划分有一定偏差,也不会影响本文研究得出的主要结论。

2 试点研究二:汉字构词歧义

2.1 语料与方法

采用试点研究一(上述)获得的汉语新闻语料(2005 年 2—3 月获得的网上新闻题目和正文)——Word 文档,给每行编号,并重新排版使每行的字符数(见下面)不超过 25 个。全部语料总共有大约 7000 行。

根据笔者以前所述方法[2],在整个语料中从头到尾的等距随机抽选 1/8 的行,从所抽选的每一行随机抽选 1/25 的字符。本文所谓的字符,原则上指的是用 Word 软件统计为一个“字数”的符号,包括汉字、阿拉伯数字等,但不包括标点符号。

对于如上抽到的每一个汉字,判断它是否有以下构词歧义。

Ⅰ类歧义:所抽汉字为一个多字词语(定义见下面)中的一个字,但它又与其前面或后面的一个或多个字构成另一个非原文含义的多字词语。

Ⅱ类歧义:所抽字为单字词语(定义见下面),但它又与其前面或后面的一个或多个字构成另一个非原文含义的多字词语。

Ⅲ类歧义:所抽字属于多字词语或单字词语,但它又可与其前面或后面的字构成非原文含义的词语或词组(非本文定义的多字词语,也未必被词典收录)。例如字段“任期内面临的问题”(表 3)里的所抽字“面”,构成原文含义的多字词语“面临”,但又可构成或理解成非原文含义的词语“内面”。“内面”这个词,词典(下面)里没有单独列出来,因此不属于本文定义的多字词语。

Ⅳ类歧义:所抽字邻近一个专有名词或为专有名词(单字或多字)的字。由于这个专有名词不常用、不熟悉或者词典(下面)里没有,因此对它的构词的划分似乎模棱两可。此类歧义实际上是第Ⅲ类歧义的一种特殊情况。

本文定义的多字词语,指的是《现代汉语词典》(商务印书馆 2005 年第 5 版)里方头括号内列出的多字条目,且这些条目的前两个或多个字不构成该词典里单独列出的任一个其他多字条目。换句话讲,本文定义的多字词语为“根词”。例如,该词典里有“物理”、“物理性质”、“物理学”等词条,只有“物理”才算本文定义的词语。又如,词典里列有“所得税”这个词条,但没有“所得”这个词条,因此后者不算本文定义的词语。如此定义是为了准确划分词语,避免实际界定字、词、词组或短语时经常出现的困难。此外,词典里未单独列出的人名、地名等专有名词(单字或多字),也算本文定义的词语。不与其前面或(和)后面的一个或多个字构成一个多字词语的单个汉字,就是单字词语。

2.1 结果

总共抽了 680 个字符,即抽选语料的字符总数估计为 13.6 万(=680×8×25)[2]。所抽字符含 7 个阿拉伯数字(例如 12%、2004)、1 个英语单词(地名)和 672 个汉字,其中 45 个汉字是用作人名、地名的汉字。

672 个所抽选的字中,有 29 个字(4.3%)有构词歧义,Ⅰ、Ⅱ、Ⅲ、Ⅳ类歧义分别占 38%、17%、24%、21%(表 3)。45 个所抽选的专有名词用字中,有 6 个(13%,见表 3 里的第 20、21、24、25、27、29 字段)有构词歧义。所有 29 个构词歧义中,有 12 个(41%,见表 3 里的第 14、17、19、20、21、22、24、25、26、27、28、29 字段)与专有名词有关。

表 3　包含构词歧义的字段及其分类与提示

歧义分类	包含构词歧义的字段	歧义提示
Ⅰ	1. 查办案件	办案？
Ⅰ	2. 大学民法学王教授	是“民法”、“法学”还是“民法学”？
Ⅰ	3. 公民享有的私人活动	有的？
Ⅰ	4. 反分裂国家法	家法？
Ⅰ	5. 交通事故	通事？
Ⅰ	6. 举世瞩目的大事	目的？
Ⅰ	7. 秘密信息	密信？
Ⅰ	8. 善后事宜	后事？
Ⅰ	9. 小组成员	组成？
Ⅰ	10. 迎新年活动	迎新？
Ⅰ	11. 执法人员	法人？
Ⅱ	12. 发言人 12 日夜宣布	日夜？
Ⅱ	13. 将人类购物分成四类	分成(提成)？
Ⅱ	14. 英国人为什么喜欢这样	人为？
Ⅱ	15. 找到王襄后人才发现的	人才？
Ⅱ	16. 这个人认为	个人？
Ⅲ	17. 车臣非法总统	法(国)总统？
Ⅲ	18. 任期内面临的问题	内面？
Ⅲ	19. 俄国家杜马主席	俄国“家杜马”？
Ⅲ	20. 期盼中国强大起来	期盼中“国强大起来”？
Ⅲ	21. 日本是在采用	本(来)是？
Ⅲ	22. 谈到台湾问题	谈“到台湾”问题？
Ⅲ	23. 这恐怕也是个税改革问题	一个“税改革”问题？
Ⅳ	24. 地处河南登封少室山林中	究竟是在哪个市/县/镇/村/山中？
Ⅳ	25. 东欧和独联体地区	是“联体”还是“和独联体”？
Ⅳ	26. 黄龙景区	“黄龙景”区？
Ⅳ	27. 借唐李贺诗句云	是“唐李”、“李贺”还是“唐李贺”？
Ⅳ	28. 卡德罗夫亲率部队	“卡德罗夫亲”率部队？
Ⅳ	29. 渥太华河北岸城市	“渥太”“华河”北岸？

注：字段中用下划线标示的字为随机抽选的字，其构词歧义的分类定义见正文。由于内容问题，字段 20、23 在原文的基础上有部分改动，但没有改变所抽字及歧义性质。

2.3　讨论

该试点研究表明，每读到含有 100 个汉字的新闻，就可能至少遇到 4 个(平均)构词歧义问题。这是一个令笔者吃惊的结果，它势必影响汉语初学者的阅读理解，尽管对熟练掌握汉语的人来讲，其在实际阅读理解过程中多不会产生那样的(构词)歧义。

构词歧义主要源于汉字连写，没有分词标示。为了避免分词歧义并便于理解，有语言学者倡导分词连写[6-7]。不过，仔细一想，分词连写不仅麻烦，还有分词困难[8]。鉴于不少歧义与人名、地名等专有名词有关(见上面的结果)，笔者赞同陈力为[6]的倡导：像我国“五四”前后至 20 世纪 50 年代期间那样，像部分古籍的整理那样，给专有名词加上专名号(下划线或波浪线等)。在现代电子新闻报道中，采用特别标示专有名词(例如用不同字体)的办法，应该并不难，也许值得一试。

关于构词歧义的定义，第Ⅰ、Ⅱ类歧义有严格定义，可以准确判断，但第Ⅲ、Ⅳ类歧义受阅历等主观因素的影响。不论怎样，有一点比较明确，当阅读理解能力不够时，或者当词汇愈来愈多时，文中的构词歧义就可能出现更多。就是说，本文观测的 672 个字中，笔者发现其中 29 个有构词歧义(表 3)，可能其中还有一些字有构词歧义，但笔者没有注意到。

计算机分词歧义一般分为交集型(占绝大多数)和组合型两种[9-12]，分别与本文定义的Ⅰ类和Ⅱ类歧义近似，但不完全相同。本文定义的是单字的构词歧义，而计算机一般定义的是多个字的歧义字段或词素段[9-13]。例如，“他的确切意图是什么”中的“的确切”是交集型歧义字段[11]，但根据本文的定义，其中的“的”字构成Ⅱ类歧义，“确”字构成Ⅰ类歧义，“切”字不构成歧义。又如，“该研究所得到的奖金很多”中的“研究所”为组合型歧义字段[11]，但根据本文的定义，其中的“所”字构成Ⅲ类歧义，但“研”或“究”字不构成歧义。笔者猜测，定义歧义字段也许适于计算机分析，但从歧义字段的多少我们不能准确了解语料中歧义发生的频率。例如，在包含 12.5 万字的新闻语料中有交集型歧义字段 504 个[14]，这并不能说明语料里有多大百分比的字或词有分词(构词)歧义。

不论是人工分词还是计算机分词，都必需基于一定的词典或词库，这是分词的依据和标准。本文的多字词语定义仅基于一本《现代汉语词典》(人名、地名除外)，有局限性，因此本文界定的Ⅰ、Ⅱ类构词歧义也有局限性。不过，本文定义的Ⅲ、Ⅳ类歧义，可以弥补Ⅰ、Ⅱ类歧义定义的局限。

3　未来相关研究思考：抽样相关问题

书面语由一系列图案、符号构成，它们先构成英

语字母、汉字等,然后构成基本结构与功能单位——单词,再构成词组、句子、段落、篇目。一个语料(在一定范围内收集的语言资料)里,有或用了什么样的字符词句,我们可以进行完全分析:研究其中所有的字符词句。当然,这样的研究可能非常有限。更多的时候,我们可能只会,实际上也只需,从语料中随机抽选部分字符词句出来进行定量研究。这就是抽样估计,彰显统计学的价值。

研究什么决定抽选什么。如果我们想研究语言运用的历史变迁,我们需要能代表不同历史时期的语料;如果我们想比较不同体裁的语言,我们需要能代表不同体裁的语料。如果我们想分析短篇笑话的幽默点,我们需要按篇目抽选笑话,然后分类归纳每篇所选笑话的幽默所在。如果我们想研究作者倾向于运用多少笔画[2]、什么发音的汉字,或者常用什么词汇、多少词汇[3],我们需要先随机抽选一定语料里的汉字、词种,然后一一分析所选汉字的笔画、发音,所选词种的性质功能、使用频率。如果我们想研究某个语料里某个标点(例如本文的试点研究一)、词的用途,某个字、词的搭配、含义(例如本文的试点研究二),我们需要先从语料里随机抽选出(或者完全找出,如果可能完全分析的话)那个标点、字或词,然后一一进行分析。

抽选什么要符合抽样原则——均匀抽样,即样本元素要从总体中随机抽选,或者说总体里每个所测元素都有相同的机会被抽选[1]。语言结构的样本元素并不复杂,就是不同层面而言的、散在分布的图案符号,例如标点符号,英语单词内的字母,汉字里的笔画,由字母笔画构成的单词,由单词构成的词组,由词组构成的句子。与生物组织结构一样,这种图案符号的随机抽样,最简单、有效、实用的办法就是等距随机抽样[1-3]。笔者以前的相关研究[2-3]以及本文的试点研究二,采用了严格的等距随机抽样。但试点研究一中引号的那种“等距”(从页数来讲)随机抽样,不是严格的引号的等距抽样,因为不同页里的引号数不同。这样按页等距抽样的结果是:引号多的页,在整个样本中的构成权重被相应的减少;而引号少的页,在整个样本中的构成权重被相应的增加。严格的引号的等距抽样可以这样进行:①先等距随机抽选页,然后观测所选页内所有的引号(前引号)。②先等距随机抽选页,然后等距随机抽选所选页内的引号(不同页的抽样间距要相同)。或者,③从整个语料等距随机抽选引号,例如从头到尾抽选整个语料中的第 3、第 23、第 43……个引号。当然,如果你不相信页内的引号数与引号用途有关,那么像本文试点研究一那样从更多页随机抽选引号的简便方法,就应该是可以接受的。

研究字种(字汇)或词种(词汇)数量与字符总数的回归关系时,我们可采用不均匀的抽样——越靠前的语料采用越小的抽样间距来进行等距抽样,以增加靠前语料内的字汇、词汇、字符总数估计的准确性[3]。但这样抽样所得的字种、词种的总的特征(例如所选字种的笔画数、词种的词性),不能无偏反映整个语料里所有字种、词种的总的特征。(这与上面引号抽样中所涉及的问题一样。)不过,这样的抽样偏差可以校正[1]。假设采用了 2 种不同抽样间距 150 和 300(即分别等距抽选 1/150 和 1/300)来抽样,那么采用抽样间距 150 所得的子样本应乘以权重系数 1,而采用抽样间距 300 所得的子样本应乘以权重系数 2。

抽选多少,即样本含量(抽选的样本元素的数量)要多大,取决于你能容忍的抽样误差。体视学中的抽样误差常用误差系数表示;把一个等距样本等分成 2 个子样本,你就可实际估计一下抽样误差[1-2]。合适的样本含量,我们也可这样一般思考。假设所测语言符号的特征(例如引号的用途)可以归为 10 个类别,那么每类抽选平均 10 个(即总共抽选 100 个)所测语言符号也许就够了,因为如果某类有多一个或少一个的误差,那平均只有 10%的高估或低估。如果你想把这个 10%降为 5%,那就总共抽选 200。如果某类符号很少,即使总共抽选 200 你也只能抽到一二个那类符号,那对于这类符号,如果它很重要的话,你可单独采用较小的抽样间距来抽选更多。(注意,对于这样单独抽选的符号,如要用其反映整个语料的情况,要乘以一个较小的权重系数,见上一段。)更有甚者,如果拟分析的符号特征(不论该有不该有,也不论正确不正确),整个语料里都没有,那就说明那种拟分析特征很少,甚至本来就没有,或者说明所用语料收集的范围或数量不够。

总而言之,随机样本可无偏反映其来源的总体。

这句话有一个推论限制，那就是样本只能反映我们所观测的总体。例如，本文的语料来自十多年前某些网页上的新闻，因此结果只能反映那个时期那种网页上的新闻语言。此外，这句话也有一个前提，那就是样本特征的观测没有系统误差或过失误差。如果我们对样本的观测有问题，例如对引号用途的划分、构词歧义的定义有问题，那我们的观测结果对总体的反映也有问题。

就应用语言的形态与功能研究而言，在有计算机的时代抽样研究还有用吗？借助计算机，我们可以建设并扩大语料库，我们也可以完全研究（非抽样研究）整个语料库。前者的确是计算机的重大优点，但后者未必非常可取，尤其是在创新研究方面。就是说，当我们还不能很好的指示计算机去怎么准确的分析测量语言形态与功能的时候，计算机就不能做好这方面的创新工作。抽样研究，只要能很好的分析样本——这是研究人员或人工分析的关键优势，就能很好、甚至更好、更有效的研究总体。

从生物组织抽选的结构，一般是“死”的，我们可以观测其形态特征，但难以分析其“活”的功能状态。而从语料中抽选的结构，是“活”的，不仅可分析其形态，也可分析其功能（用法）。这就是为什么笔者认为，抽样估计在定量语言学研究中具有较大的潜在应用价值。

后记

（1）本文的研究以及笔者前期的研究[2-3]所用抽样方法——等距抽样，也称系统抽样或机械抽样，是一般统计学的基本抽样方法之一，并非只用于体视学，也不是体视学所独有，只是常用于体视学研究，而且体视学把等距抽样用到了极致——甚至用其抽样间距来估计粒子总数[1-3]。

（2）笔者不是职业语言学者，对语言相关领域的知识与研究的了解和理解有限，但笔者熟悉体视学抽样并认为它可用于进行一些有趣或重要的语言相关研究，因此尝试进行了相关研究，希望能起到抛砖引玉的作用。

参考文献（References）

[1] 杨正伟．生物组织形态定量研究基本工具：实用体视学方法[M]．北京：科学出版社，2012.
Yang Zhengwei. Essential Tools for Morphometric Studies of Biological Tissues: Practical Stereological Methods [M]. Beijing: Science Press, 2012.

[2] 杨正伟，秦诗芸．体视学分合法在中英文书籍字、词数估计中的运用研究[J]．中国体视学与图像分析，2009，14(3)：271-278.
Yang Zhengwei, Qin Shiyun. Use of the stereological fractionator in the number estimation of characters and words in Chinese and English books [J]. Chinese Journal of Stereology and Image Analysis, 2009, 14(3): 271-278.

[3] 杨正伟，贺显利，许薇，等．不同字符数量的体视学抽样估计：方法介绍与试点研究[J]．中国体视学与图像分析，2017，22(1)：87-90.
Yang Zhengwei, He Xianli, Xu Wei, et al. Stereological sampling estimation of the number of different characters: Methodology introduction and pilot study [J]. Chinese Journal of Stereology and Image Analysis, 2017, 22(1): 87-90.

[4] 北京大学中文系现代汉语教研室．现代汉语（重排本）[M]．北京：商务印书馆，2004.
Department of Modern Chinese Teaching and Research, Faculty of Chinese Language and Literature, Beijing University. Modern Chinese (Reprint) [M]. Beijing: Commercial Press, 2004.

[5] Swan M. Practical English Usage [M].Oxford: Oxford University Press, 1995. 471.

[6] 陈力为．汉语书面语的分词问题：一个有关全民的信息化问题[J]．中文信息学报，1996，10(1)：11-13.
Chen Liwei. Word segmentation of written language, a question of all people's information [J].Journal of Chinese Information Processing, 1996, 10(1): 11-13.

[7] 张小衡．也谈汉语书面语的分词问题——分词连写十大好处[J]．中文信息学报，1998，12(3)：57-63.
Zhang Xiaoheng. Written Chinese word-segmentation revisited: ten advantages of word-segmented writing [J].Journal of Chinese Information Processing, 1998, 12(3): 57-63.

[8] 熊文新．汉语真需要词间空格吗？对汉语分词连写献疑[J]．语言科学，2014，13(6)：655-669.

Xiong Wenxin. Does Chinese need spaces between words [J]. Linguistic Sciences, 2014, 13(6): 655-669.

[9] 曹倩,丁艳,王超,等. 汉语自动分词研究及其在信息检索中的应用[J]. 计算机应用研究,2004,5: 71-74, 91.
Cao Qian, Ding Yan, Wang Chao, et al. A survey of Chinese word segmentation and the application in information retrieval [J]. Application Research of Computers, 2004, 5: 71-74, 91.

[10] 陈其晖,应志伟,柴佩琪. 基于歧义二叉树的汉语分词方法[J]. 计算机辅助工程,1999,4: 12-17.
Chen Qihui, Ying Zhiwei, Cai Peiqi. Ambiguity binary tree based Chinese word segment [J]. Computer Aided Engineering, 1999, 4: 12-17.

[11] 周昌乐,秦莉娟. 一种采用基于语境松弛算法的汉语分词排歧方法[J]. 厦门大学学报(自然科学版),2002,41(6): 711-714.
Zhou Changle, Qin Lijuan. A disambiguation method for segmenting Chinese words by using relaxation algorithm based on context [J]. Journal of Xiamen University (Natural Science), 2002, 41(6): 711-714.

[12] 张莉莉,冯燕. 基于语料库的汉语自动分词错误类型分析[J]. 华中师范大学研究生学报,2017,24(1): 93-97.
Zhang Lili, Feng Yan. A corpus-based analysis of the errors in Chinese word segmentation [J]. Central China Normal University Journal of Postgraduates, 2017, 24(1): 93-97.

[13] 徐秉铮,贺前华. 汉语自动分词歧义及处理策略[J]. 中文信息,1992,1: 17-20.
Xu Bingzheng, He Qianhua. Ambiguity of automatic Chinese word-segmentation and its management strategy [J]. Chinese Information, 1992, 1: 17-20.

[14] 孙茂松,黄昌宁,邹嘉彦,等. 利用汉字二元语法关系解决汉语自动分词中的交集型歧义[J]. 计算机研究与发展,1997,34(5): 332-339.
Sun Maosong, Huang Changning, Zou Jiayan, et al. Using character bigram for ambiguity resolution in Chinese word segmentation [J]. Computer Research and Development, 1997, 34(5): 332-339.