文章编号：1007 - 1482（2009）03 - 0271 - 08 ·论著·

体视学分合法在中英文书籍字、词数估计中的运用研究

杨正伟[1]， 秦诗芸[2]

（1. 川北医学院 形态定量研究室， 南充 637007；
2. 川北医学院 外国语言文化系 2004 级 1 班， 南充 637007）

摘 要： 本研究利用现代英汉（7 套书）和汉英（3 套书）对照语料，根据体视学分合法估计英语和汉语语料中的英语词以及汉语字、词的数量，以尝试对比这两种语言的基本结构特征。结果表明，在英汉、汉英对照语料中，1 个英语词分别平均表达或传递了 1.47 ~ 1.72、1.04 ~ 1.41 个汉字和 1.02 ~ 1.23、0.78 ~ 0.90 个汉语词的含义或信息。这提示，英译汉时译者倾向于使用相对较多的汉语字与词，而汉译英时则倾向于使用相对较多的英语词。

关键词： 英语；汉语；字；词；数量；体视学；分合法
中图分类号： H087；H13；H313；N3 **文献标识码：** A

Use of the stereological fractionator in the number estimation of characters and words in Chinese and English books

YANG Zhengwei[1]， QIN Shiyun[2]

（1. Morphometric Research Laboratory, North Sichuan Medical College, Nanchong 637007, China;
2. Class 1 Grade 2004, Foreign Language and Culture Department,
North Sichuan Medical College, Nanchong 637007, China）

Abstract: The study aimed to tentatively compare the basic structural characteristics of the current English and Chinese languages by utilizing English-Chinese (7 book sets) and Chinese-English (3 book sets) materials and estimating the numbers of English words and Chinese characters and words in the English and Chinese materials with the stereological fractionator method. The results showed that one English word, in the English-Chinese and Chinese-English materials respectively, expressed or conveyed the meaning or information of 1.47 ~ 1.72 and 1.04 ~ 1.41 Chinese characters, and 1.02 ~ 1.23 and 0.78 ~ 0.90 Chinese words on average. This suggests that relatively more Chinese characters and words tend to be used by translators in translating English into Chinese, whereas relatively more English words tend to be used in the translation of Chinese into English.

Key words: English; Chinese; characters; words; number; stereology; fractionator

就使用的人数而言，英语和汉语无疑是当今世界上最常用的两种语言。这两种语言尽管在起源和表达符号等方面很不相同，但也有一定的相似性：它们都基本上属于孤立语（isolating or root language），其基本结构与功能单位都是词，词序的改变是语法的基础[1]。

收稿日期：2009-06-23
作者简介：杨正伟，教授，研究方向：生物组织结构的体视学定量研究。E-mail：zwyang@ nsmc. edu. cn

英语的词(单词)由一个或多个字母构成;汉语的词由一个或多个字构成,一个字又由一个或多个笔画构成。可以说,两种语言都可表达相同的内容。不过,由于两种语言(书面语)的词汇和语法不完全对应,表达同一内容或传递同样信息所需的词和标点等的数量自然会有不同,有时需要更多的英语词,有时需要更多的汉语词。总的来讲(当所表达的内容足够多时),一个英语词平均可表达或传递多少汉语字或词的信息呢?这是两种语言结构对比的一个基本问题,也是一个有趣的问题,但笔者尚未见到这方面的研究报告。要回答这个问题,最好利用表达同样内容、传递同样信息量的英语及其对应汉语语料,并对比两种语料中的总词数等。如何获得一本纸质语料中的总词数?如果要一个一个词的整本计数,想必没有人愿意去做这样的研究。有没有一种科学的抽样计数方法?正好,体视学中有一种抽样估计粒子(细胞等散在可数的结构)数的方法——分合法(fractionator)[2]。这种方法实质上是通过一系列等距随机抽样(例如器官组织块的等距随机抽样、组织块连续切片的等距随机抽样、切片内测试视野的等距随机抽样),最后从器官抽取少量(若干分之一)体积的组织来计数其中的粒子数,然后根据所抽组织的比例以及从所抽组织计数的粒子数估计器官内的粒子总数[2]。笔者认为,我们同样可通过书中页、页中行、行中词等的一系列等距随机抽样,最后从书中抽取少量的词等来计数以估计书内词等的总数。因此,本文选用多套英汉和汉英互译或对照小说、教材等语料(纸质书籍),借用分合法估计英语及其对应汉语语料中字、词、标点、字母、笔画的数量,以尝试对比这两种语言的结构特征。

1 语料与方法

语料

(1)小说《飘》,其英文版由当代中国出版社出版(2004 年第 2 版),汉译版由人民文学出版社出版(1990 年第 1 版,2004 年第 4 次印刷)。汉译者:戴侃,李野光,庄绎传。

(2)小说《简·爱》,其英文版由当代中国出版社出版(2004 年第 2 版),汉译版由人民文学出版社出版(1990 年第 1 版,2004 年第 5 次印刷)。汉译者:吴钧燮。

(3)英汉对照小说(改编的简易读物)《螺丝在拧紧》,由航空工业出版社出版(2005 年第 1 版)。汉译者:毛荣贵,姜淑萱。

(4)英汉对照教材《新概念英语》第 1 ~ 4 册,由外语教学与研究出版社与朗文出版亚洲有限公司出版(1997 年第 1 版,2005 ~ 2006 年重印)。编者:L. G. Alexander,何其莘。

(5)汉英对照小说《围城》,由人民文学出版社出版(2003 年第 1 版,2004 年第 3 次印刷。)英译者:珍妮·凯利,茅国权。

(6)汉英对照公文《中国政府白皮书(3),2000 –2001》,由外文出版社出版(2002 年第 1 版)。编者:中华人民共和国国务院新闻办公室。

(7)汉英对照散文《英译中国现代散文选,第 2 辑》,由上海外语教育出版社出版(2003 年第 1 版,2005 年第 4 次印刷)。英译者:张培基。

标点、字、词、字母、笔画数的抽样估计:语料 1、2、3、5、6 的正文(不包括注释内容)内标点、字、词、字母、笔画的总数等,采用下述方法进行抽样估计。

标点、汉字、英语词数的抽样估计:首先从英语或汉语语料中等距随机抽取 $1/n_p$ 页,然后从所抽取的各页中等距随机抽取 $1/n_l$ 行,再从所抽取的各行内抽取所有标点符号,并等距随机抽取 $1/n_w$ 个“字词”。n_p、n_l、n_w 均为正整数,分别表示页、行和行内“字词”的抽样间隔。语料里标点符号的总的抽样间隔 n 为($n_p \cdot n_l$),“字词”的总的抽样间隔 n 为($n_p \cdot n_l \cdot n_w$)。假如从英语或汉语语料里最后抽得的标点或“字词”总数为 Q,那么语料里的标点或“字词”总数 N 的无偏估计为:

$$N = n \cdot Q \qquad (1)$$

上述所谓的“字词”,本文指的是英语词、汉字、阿拉伯数字等特殊符号,不包括标点符号。1 个“字词”,用 Microsoft 的 Word 软件来统计的“字数”原则上为 1。例如,英语语料中的“I'll、sculptor's、U. N. 、thirty-seven”各算 1 个英语词(共 4 个),汉语语料中的“三十七”算 3 个汉字,英语或汉语语料中“37、LFZ312G”各算 1 个“字词”(共 2 个)。对于跨行的英语词,采用算头不算尾的计数原则。例如“to-morrow”,其在上一行中的部分“to-”算作 1 个英语词,即该单词“tomorrow”,而其在下一行的部分“morrow”不算作任何“字词”。

页、行和行内的抽样间隔事先任意大致确定,以使最终能抽到的“字词”总数大致不少于 200。对于上述语料,本文实际所用的页、行和行内的抽样间隔分别为 1 ~ 4、15 ~ 40、15 ~ 29,实际最终从英语或汉语语料 3 中抽到 130 ~ 159 个标点,61 ~ 67 个“字词”;从英语或汉语语料 1、2、5 或 6 中抽到 433 ~ 1 022个标点,215 ~ 349 个“字词”。

下面以英语版《飘》为例说明具体抽样估计过程。该书共 1231 页,每页内的行数不超过 31,每行内的“字词”数很少超过 15。我们事先确定让页、行和行内的抽样间隔分别为 4、31、15。首先在 1 ~ 4 之间随机确定一个均匀随机正整数:用 Microsoft 的 Excel 软件(电子表格),键入公式“ = RAND( ) * 4 + 0. 5”(式中的“RAND( )”是 0 ~ 1 之间的均匀随机数);实际确定为 1(小数点四舍五入),因此,以 4 为抽样间隔等距随机抽取该书中的第 1、5、9、13、…、1 229 页。所抽取的各页内拟抽取的 1 行(由于每页的行数不超过 31,每页内最多只能抽到 1 行)的序号,用 Excel 软件的公式“ = RAND( ) * 31 + 0. 5”确定(小数点四舍五入)。所抽取的各行内拟抽取的“字词”的序号,用 Excel 软件的公式“ = RAND( ) * 15 + 0. 5”确定(小数点四舍五入);对于“字词”数超过 15 的所抽行,再确定一个等距随机序号:在如上确定的该行的随机序号 r 的基础上加 15,即再抽取该行内的第(r + 15)个“字词”(能抽到就抽,不能抽到就不抽)。最后,实际从该书共抽得 308 页,277 行;这 277 行内的标点总数为 474,其中起点断作用的点号(句号、问号、叹号、逗号、分号、冒号)数为 320;从这 277 行内共抽得 215 个“字词”,其中 214 个为英语词(其中 7 个单词“an’you、B. 、hadn’t、I’ll、It’s、It’s、O’Hara”有词内或词末标点),1 个为阿拉伯数字“1864”。因此该书中的标点符号、点号、“字词”、英语词总数的无偏估计分别为 58 776、39 680、399 900、398 040。

上述抽样估计方法即体视学中常用的分合法[2],其原理非常简单,相当于从总体中等距随机(systematic random)抽取了 1/n 的量(样本量),将之乘以 n 即为总体量的无偏估计。例如,对于有序的 Y(假设为 99)个元素(1、2、3 ,…, 99),先任意确定一个正整数 n(假设为 20),然后在 1 至 20 之间随机确定 1 个正整数(假设为 3),由此确定从这 99 个元素中抽取序号为 3、23、43、63 和 83 的元素。此即等距随机抽样,n 为抽样间距,共抽到 5 个元素。因此,Y 的估计为 $20 \times 5 = 100$(根据公式 1)。假设 n 确定为 110,又假设从 1 ~ 110 间随机确定的数字为 102,那么从 Y(99)个元素中能抽到的元素个数为 0,此时,Y 的估计为 $110 \times 0 = 0$(根据公式 1)。由于从 Y 个元素中能抽到的元素个数平均(期望值)为 Y/n,根据公式 1 估计的结果正好等于 Y,因此这种抽样估计是无偏估计。显而易见,为了减少抽样误差,n 最好不要大于 Y,且 Y 最好(接近)为 n 的整倍数(例如 n 为 5、10 或 20),这样就会显著减少抽样误差。即是说,利用等距随机抽样的分合法估计,重要的是要注意通过适当的抽样设计来尽量减少抽样误差(见上述例子以及“讨论”部分)。

对上述抽样估计的误差,本文如下进行了分析:

把从某语料中等距随机抽取的所有页,分成 2 个等距子样本,即抽得的第 1、3、5、…页和抽得的第 2、4、6、…页各构成 1 个子样本。根据公式 1 从这 2 个子样本各得 1 个估计值 N_1、N_2,然后如下估计误差系数(CE, coefficient of error)[3]:

$$CE = \frac{1}{\sqrt{2}} \cdot \frac{| N_1 - N_2 |}{(N_1 + N_2)} \qquad (2)$$

该误差系数反映的是最终估计结果(2 个子样本合在一起进行估计所得的结果)的误差——估计值与真值之间的平均差异[3]。

英语词字母数、汉字笔画数的抽样估计:在上述行内“字词”的抽样过程中,每抽到一个英语单词(对于英语语料),就计数其字母数;每抽到一个汉字(对于汉语语料),就计数其笔画数。所抽英语词的字母总数除以所抽英语词的总数,即得每个英语词的平均字母数;所抽汉字的笔画总数除以所抽汉字的总数,即得每个汉字的平均笔画数。该平均值估计的误差系数根据公式 3(见下述)估计。所抽英语词的字母总数或汉字的笔画总数,乘以英语词或汉字的总的抽样间隔($n_p \cdot n_l \cdot n_w$),即分别得所测语料中的字母或笔画总数(根据公式 1 计算),该总数估计的误差系数根据公式 2 计算。

汉语词数的抽样估计:对于汉语语料,在上述行内“字词”的抽样过程中,每抽到一个汉字,就按原文含义确定该字是否与其前面或(和)后面的一个或多个字构成一个没有标点符号隔开的多字词。如不构成一个多字词,该字就是一个单字词。如构

成一个多字词,就进一步确定该词是否是首字词,即所抽字为该词的第一个字(词首字)。如所抽字构成首字词(单字词也算作首字词),就抽选该词,反之不抽选。这样抽选词,相当于抽选各行内的单字词的字以及多字词的首字。把如此抽选的首字词的数量总和起来,然后根据公式 1(该式中的抽样间隔 $n = n_p \cdot n_l \cdot n_w$)无偏估计语料里的汉语词总数。

如上通过抽选词首字来抽选词的方法,是本研究的"窍门",犹如利用体视框来抽选粒子(一个一个的词或"词串"实际上相当于一个一个的粒子)的"头"部[4]。由于每个词有并且只有 1 个词首字,因此根据公式 1 的词数估计是无偏估计。例如,假设在 X 个字中有 Y 个词,即有 Y 个词首字,那么,在等距随机抽选的 $1/n$ 个字中,词首字个数的平均(期望值)为 $[(X/n) \cdot (Y/X)] = Y/n$,将之乘以 n(根据公式 1)正好得 Y。

本文所定义的多字词,主要指的是《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,如此定义的多字词为"根词"。例如,在"物"字头下,该词典里列有"物理"、"物理性质"、"物理学"等词条,而该词典在"性"字头下单独列有词条"性质",因此,"物理"是本文定义的 1 个双字词,而"物理性质"由 2 个双字词构成,"物理学"由 1 个双字词和 1 个单字词(学)构成。又如,该词典里列有"所得税"这个词条,但没有"所得"这个词条,因此"所得税"是 1 个多字词,而"所得结果"由 2 个单字词(所、得)和 1 个双字词构成。(据笔者用分合法估计,本文所用《现代汉语词典》有约 4.9 万个本文所定义的多字词。)这样定义词是为了准确划分词,以避免实际界定字、词、词组或短语时会经常出现的困难。此外,该词典里没有的人名、地名等专有名词(单字或多字)以及阿拉伯数字等特殊符号,也算本文定义的词。例如,汉语语料中的"钱钟书"、"三十七"、"37"、"LFZ312G"分别算作 1、3、1、1 个词,后 2 个词不是由汉字构成,属于本文所谓的阿拉伯数字等特殊符号。

字、词、标点的完全计数

对于语料 4(新概念英语)、7(散文选),从每本语料中等距随机抽选 9 ~ 10 篇课文或散文(英语及其对应的汉语),然后完全计数所抽课文(或课文的一部分)内的所有标点、"字词"、汉语词。"字词"、汉语词的定义如上所述。

从 4 册新概念英语各抽选了 10 课进行测量。对于第 1、2 册,测量了整个所抽课文;对于第 3、4 册,由于课文较长,仅测量了所抽各课的总行数分别不少于 6、5 的前 n(最小正整数)个段落(以英语为准),即用部分段落代表课文以减少工作量。从散文选语料抽选了 9 篇散文,对每篇所抽散文仅测量了总行数不少于 10 的前 n(最小正整数)个段落(以汉语为准),即用部分段落代表课文以减少工作量。所抽并测量的课文的内容,均针对的是正文,不包括题目、作者、注释以及对话课文(新概念英语第 1 册)左边的人名(非对话内容)。最后实际从新概念英语第 1 ~ 4 册和散文选的每篇所抽英语或汉语课文(或课文的一部分)中,分别计数到 14 ~ 49、14 ~ 71、6 ~ 38、8 ~ 37、17 ~ 64 个标点,45 ~ 220、97 ~ 338、65 ~ 418、74 ~ 482、126 ~ 409 个"字词"。

把每本语料的 n(9 ~ 10)篇课文的结果平均,得到每本语料的平均结果 $\bar{x}$,并计算标准差 SD,然后如下计算误差系数(根据常规的统计方法):

$$CE = \frac{SD}{\sqrt{n} \cdot \bar{x}} \qquad (3)$$

该误差系数既反映那本语料的平均结果估计的误差,也反映那本语料的课文间的变异(式中的"$SD/\bar{x}$"为课文间的变异系数)。

2 结果

英语单词、汉语字与词的总数

英语小说《飘》、《简爱》、《螺丝在拧紧》中的"字词"总数(英语词数与阿拉伯数字等特殊符号数之和)分别为 399 900、183 923、25 620。从英语《飘》中抽到 1 个阿拉伯数字,从英语《简爱》中抽到 2 个特殊符号——法语,分别占所抽"字词"总数的 0.47%、0.57%;从汉语《飘》、《简爱》以及英汉《螺丝在拧紧》语料中未抽到阿拉伯数字等特殊符号。从《新概念英语》第 1 ~ 4 册的英语语料、对应汉语语料中,数到的阿拉伯数字的数量分别占"字词"总数的 0 ~ 0.48%、0.19% ~ 0.80%;除第

1 册中数到 1 个特殊符号“LFZ312G”外,该 4 册语料中没有数到其他特殊符号。

汉语小说《围城》、公文《中国政府白皮书》中的“字词”总数(汉字数与阿拉伯数字等特殊符号数之和)分别为 178 416、126 765,词总数(汉语词数与阿拉伯数字等特殊符号数之和)分别为 132 300、76 950。在《围城》以及《英译中国现代散文选》的汉语及其对应英语语料里没有数到阿拉伯数字等特殊符号,而在《中国政府白皮书》的汉、英语料里数到的“字词”中,阿拉伯数字(没有数到其他特殊符号)的数量分别占 1.60%、5.11%。

利用分合法测量的语料中,除语料 3 略高以外,其他英语或汉语语料的“字词”总数、汉语词总数或标点符号总数估计的误差系数均小于 5%(见表 1)。

英语测量值与对应汉语测量值之比

7 套英汉对照语料的英语词总数与对应汉语字总数(这里词或字总数严格地讲是“字词”总数)之比为 0.58 ~ 0.68(平均 0.64),均低于 3 套汉英对照语料的该比值(0.71 ~ 0.96,平均 0.83)(表 1)。经统计学检测(假定所测英汉和汉英语料为随机样本),英汉语料的该比值显著低于汉英语料的该比值(Mann-Whitney 秩和检验:$P < 0.05$),平均低约 23%。

7 套英汉对应语料的英语词总数与对应汉语词总数之比(0.81 ~ 0.98,平均 0.88),均低于 3 套汉英对照语料的该比值(1.11 ~ 1.29,平均 1.19)(表 1),前者显著低于后者(t 检验:$P < 0.001$),平均低约 26%。

表 1　英语语料的词、标点总数与对应汉语语料的字、词、标点总数之比

	词-字数比	词-词数比	标点数比	点号数比
英汉对应语料				
飘	0.58(0.3%, 0.8%)	0.81(0.3%, 1.5%)	0.86(2.1%, 1.7%)	0.70(1.8%, 0.6%)
简爱	0.64(1.4%, 0.2%)	0.81(1.4%, 1.1%)	1.08(0.1%, 1.8%)	0.97(1.1%, 0.4%)
螺丝在拧紧	0.67(5.8%, 1.1%)	0.83(5.8%, 10.5%)	0.76(9.8%, 4.9%)	0.59(5.4%, 6.3%)
新概念英语第 1 册	0.68[4.0%]	0.89[2.9%]	1.28[4.0%]	1.01[3.0%]
新概念英语第 2 册	0.65[3.3%]	0.92[2.1%]	0.94[5.2%]	0.86[4.5%]
新概念英语第 3 册	0.66[3.2%]	0.98[3.3%]	0.90[5.0%]	0.87[4.4%]
新概念英语第 4 册	0.62[3.5%]	0.92[3.2%]	0.88[5.4%]	0.83[5.2%]
汉英对照语料				
围城	0.96(0.5%, 1.8%)	1.29(0.5%, 2.0%)	1.28(0.6%, 0.5%)	0.92(1.0%, 1.8%)
中国政府白皮书	0.71(2.8%, 2.9%)	1.17(2.8%, 3.0%)	0.79(4.1%, 4.5%)	0.64(0.2%, 3.4%)
英译中国现代散文选	0.83[4.9%]	1.11[4.5%]	0.98[11.2%]	0.81[9.0%]

圆括号内的 2 个百分数分别指英语语料的词或标点总数估计的误差系数及其对应汉语语料的字、词或标点总数估计的误差系数(根据文中公式 2 计算),方括号内的 1 个百分数指的是英-汉比值估计的误差系数(根据文中公式 3 计算)。

《新概念英语》各课文的英语词数与对应汉语字或词数之比(表 1),4 册之间没有显著性差异(单向方差分析:$P > 0.05$)。

单独就英语与对应汉语的标点数比或点号数比(表 1)而言,3 套英汉对照语料(新概念英语除外)与 3 套汉英对照语料相比较,两者之间的变异大,看不出明显的规律。不过,9 套英汉和汉英对照语料(新概念英语第 1 册除外)的点号数比皆小于 1(表 1),平均为 0.80。标点数比与点号数比这两个比值(表 1)相比较,不论是在英汉还是汉英对应语料中,前者均高于后者。10 套英汉和汉英对照语料中,汉语语料中的点号总数占标点符号总数的百分比平均为 86.0%(根据公式 3 计算的误差系数为 2.1%),显著高于英语语料中的点号百分比(平均 73.1%,误差系数 4.1%),平均高约 19%(配对 t 检验:$P < 0.001$)。英语里特有(汉语里没有)的表示省略、所有格和连接的词内非点号标点符号(例如 I'll 和 sculptor's 中的撇号,U. N. 中的圆点,thirty-seven 中的连字号),在英语语料的标点符号总数中平均(10 套语料的平均)占 13.9%(误差系数 15.0%)。

字母数与笔画数

3 套英汉和 2 套汉英对应语料中,简易读物《螺丝在拧紧》的英语词的平均字母数和对应汉字的平均笔画数均最低(表 2)。3 套英汉对应语料中的平均字母数或平均笔画数,均低于 2 套汉英对应语料中的结果(表 2),不过这个差异没有统计学意义(t 检验:$P>0.05$)。

3 套英汉对应语料的英语词字母总数与对应汉字笔画总数之比,均显著低于 2 套汉英对应语料的结果(t 检验:$P<0.05$),平均低约 33%(表 2)。

表 2 英语单词字母数与汉字笔画数

	每个英语单词的平均字母数	每个汉字的平均笔画数	字母-笔画总数比
英汉对应语料			
飘	4.29 [3.4%]	7.03 [2.7%]	0.35 (2.2% , 1.8%)
简爱	4.16 [2.9%]	6.89 [2.2%]	0.38 (2.3% , 0.9%)
螺丝在拧紧	3.84 [6.5%]	6.66 [5.6%]	0.39 (4.5% , 1.8%)
汉英对照语料			
围城	4.66 [3.1%]	7.24 [2.9%]	0.61 (3.3% , 3.6%)
中国政府白皮书	5.28 [3.1%]	7.10 [2.2%]	0.51 (0.8% , 1.3%)

方括号内的 1 个百分数指的是平均值估计的误差系数(根据文中公式 3 计算),圆括号内的 2 个百分数分别指英语单词字母总数及其对应汉语汉字笔画总数估计的误差系数(根据文中公式 2 计算)。

3 讨论

英语的 1 个单词就是 1 个词,而汉语的 1 个词由 1 个或多个汉字构成,因此 1 个英语词可能总的会超过 1 个汉字所包含的含义或传递的信息。该研究显示,英汉和汉英对应语料的英语词数与对应汉语字数之比为 0.58 ~ 0.96(表 1)。因此,总的来讲,1 个汉字包含的信息对等于 0.58 ~ 0.96 个英语词的,换言之,1 个英语词平均传递了 1.04 ~ 1.72 个汉字的信息。该研究也显示,英语词数与对应汉语词数之比为 0.81 ~ 1.29(表 1),即 1 个英语词平均传递了 0.78 ~ 1.23 个汉语词的信息。有趣的是,这种信息对比受翻译过程的影响:1 个英语词(原文)对等的汉译字、词数(分别为 1.47 ~ 1.72、1.02 ~ 1.23 个)相对较多,而 1 个翻译的英语词对等的汉语字、词(原文)数(分别为 1.04 ~ 1.41、0.78 ~ 0.90 个)相对较少。这说明,英译汉时译者使用了相对较多的汉语字与词,而汉译英时使用了相对较多的英语词。支持该结论的另一组数据是,英汉对照语料中的英语词的字母总数与对应汉语字的笔画总数之比显著小于汉英对照语料中的该比值(表 2)。因此,如果要问表达同样的内容时英语与汉语相比哪种语言更简练,本研究的结果显示,就两种语言的基本结构与功能单位——词的使用频数而言,英汉对应语料中英语较简练,而汉英对照语料中汉语较简练;如果把英译汉和汉译英平均起来考虑,两种语言几乎一样简练,因为英汉、汉英对应语料中的英语词数与对应汉语词数之比(分别为 0.88、1.19)的平均约为 1(表 1)。

尽管 4 册《新概念英语》教材在英语阅读理解的难易度上有明显的梯次差异,但其英汉对应的词-字数比或词-词数比(表 1)在 4 册之间没有显著性差异,即译成汉语时使用的汉语字与词的频数并没有随英语难易度的变化而变化。这也许提示,就该语料涉及的内容而言,可表达这些内容的汉语词汇足够的丰富,对较难内容的汉译表达不需要使用更多的词。

该研究表明,汉语中使用点号的频数比英语平均多约 19%。但《新概念英语》第 1 册的英语和对应汉语中的点号使用频数非常接近(表 1),这显然是因为该书的课文几乎全为日常生活口语,句子简单,其对应的汉译也简单。汉语中点号的使用较多,可能是由汉语本身的语言特征所决定的。例如,汉语语句中的词与词之间无间隔,容易导致分词歧义,使用更多的点号有助于避免歧义。又如,英语中的限定性定语从句可放在所修饰词的后面,

从句与其修饰词之间不加标点符号;如用汉语来表达或翻译,就可能要使用点号以清晰表达或避免歧义。不过,英语中有表示省略、所有格、连接等的词内或词末符号,而汉语没有。该研究显示,这种特有符号在英语标点符号总数中平均占 14%,这使汉语中非点号的标点符号的使用相对较少。这一"多"—"少"的结果是,英语与对应汉语中的标点符号总数相差不大(表 1)。

英语用词有一个规律,那就是字母愈少的词使用的频数趋于愈大[1]。汉语也有一个类似规律:笔画愈少的字使用的频数趋于愈大[5]。简易读物的词汇较少,使用的常用词汇较多,因此,简易读物的英语单词的平均字母数或汉字的平均笔画数可能会少于原版读物的,这得到了本研究结果的支持:简易对照读物《螺丝在拧紧》中的英语单词的平均字母数及汉字的平均笔画数均最小(表 2)。该研究也显示,英汉对应语料的平均字母数和平均笔画数皆小于汉英对应语料的结果(表 2)。这可能说明,本文所选用的英汉语料与本文所选用的汉英语料相比,前者使用了相对较多的常用词汇;或者,本文所选英汉语料的汉译者使用了相对较多的常用汉字,而汉英语料的英译者使用了相对较少的常用英语单词。

本研究的一个新颖之处是借用了既简单又无偏的方法——分合法来估计书籍(纸质语料)中字、词等的数量。分合法实质上是借用等距抽样及其抽样间隔进行总量估计的统计方法。这种方法于 20 世纪 30 年代提出,但一直被忽视,直到 80 年代被 H. J. Gundersen(1986)"挖掘"出来后才得到广泛的运用,主要是与体视学中的体视框技术联合运用,用于估计生物组织内粒子(例如人小脑内蒲肯野细胞)的数量,其至关重要的优点是:结合分合法的粒子数估计不受组织处理过程中结构体积变化(皱缩或膨胀)的影响[2,4,6]。把字、词、标点等语言符号当作粒子,本文首次借用分合法对比研究了英汉两种语言的结构特征。与生物细胞不同的是,"语言粒子"不需要借助切片来观察,我们看到的都是一个个完整的粒子。"语言粒子"相当于 2 维空间(平面)内的 2 维粒子;我们也可把它看作 3 维空间内的 2 维或 3 维粒子,甚至看作 1 维空间(线)内的 0 维点。这并不影响本文的研究。就汉语词数估计而言,本文的另一个新颖点是通过计数词首字来计数词。这种方法参考了用体视框计数粒子的体视学方法。

分合法估计的误差(抽样误差)大小,取决于等距抽样所得样本间的变异;就书而言,取决于所抽页数、页内行数以及行内字词数的变异。通过选择适当的抽样间隔,可有效控制抽样误差,见"标点、汉字、英语词数的抽样估计"方法部分的例子。再举例讲,假如某书刚好有 10 页,每页刚好有 10 字,把页的抽样间隔定为 5 比定为 4 要好,因为以 5 为抽样间隔的估计结果(假设每页内不再进行抽样)均为 100 字(抽样误差为 0),而以 4 为抽样间隔的估计结果,有 2/4 的可能是 120 字,有 2/4 的可能是 80 字(即有抽样误差;但所有可能估计结果的平均刚好为 100 字——真值,所以这种估计是无偏的)。又如,某书每页的行数都不超过 31,且大多数页的行数都是 31;把每页内的行的抽样间隔定为 15、30 或 31,比定为 7、20 或 40 要好,因为后者使从每页中抽到的行数的变异增大。从本文的误差分析来看,当最终抽(数)到的字、词的数量超过 200 时,都得到了满意、可靠的字、词总数估计结果:误差系数小于 5%。

对于《新概念英语》和《英译中国现代散文选》语料中的所抽选课文,为获得其字词总数等,本研究采用的是精确的完全计数方法,而没有采用有抽样误差的抽样计数方法。这是因为所抽课文的字词数少,容易数完整个所抽课文,没有必要进行抽样计数。选用由不同作者撰写的短篇课文构成的语料来研究也非常重要,因为用它可研究语言特征的另一个重要方面:课文或作者之间的变异情况。该研究显示,尽管所测各课的字词数并不多(在 45 ~482 之间),但其英汉对应的词-字数比或词-词数比(表 1)的课文之间的变异并不是很大:各套语料的词-字数比或词-词数比的变异系数在 7% ~15% 之间。

本研究所用语料的量与题材均非常有限。本文所定义的汉语多字词,也主要是一本《现代汉语词典》中所列的多字词。此外,本文采用的是手工测量方法,虽简单易行但较麻烦。如有英汉和汉英对应语料的电子文档,就可能借助普通或专门的电脑软件更快更好的完全计数整个语料中的标点符号、英语单词及其字母、汉字及其笔画,甚至汉语词。但对于纸质语料,像本文这样进行抽样估计(包括结合课文抽样的完全计数),不失为一个科学的便利选择;对于汉语词数估计而言,手工的抽

样估计还可避免分词歧义问题。总之,就语言研究而言,本研究只是一项尝试性的初步研究;不过,该项尝试至少对体视学分合法的介绍具有教学意义。

参考文献:

[1] Crystal D. The Cambridge Encyclopedia of Language [M]. 2nd ed. Cambridge: Cambridge University Press, 1997.

[2] Gundersen H J. Stereology of arbitrary particles. A review of unbiased number and size estimators and the presentation of some new ones, in memory of William R. Thompson [J]. J Microsc, 1986, 143(Pt - 1): 3 - 45.

[3] Yang Zhengwei, Zhang Rendong, Wen Xiaohong, et al. Caveat on the error analysis for stereological estimates [J]. Image Anal Stereol, 2000, 19(1): 9 - 13.

[4] 杨正伟. 生物体视学新工具——光学体视框 [J]. 中国体视学与图像分析, 1998, 3(1): 50 - 54.

[5] 苏培成. 现代汉字纲要 [M]. 第 2 版, 增订本. 北京: 北京大学出版社, 2001.

[6] 唐勇, 杨正伟, 崔成虎. 分合法研究成年男性小脑蒲肯野氏细胞数目 [J]. 川北医学院学报, 1992, 7(3): 1 - 4.

·动态与信息·

CNKI 数字图书馆全文数据库中国学术团体专用卡

学会现有少量的 CNKI 数字图书馆全文数据库中国学术团体专用卡,如有会员需要可以到中国体视学学会(清华大学工物馆 113-A)免费领取,面值 50 元。

学术团体专用卡具体使用程序:

进入学会网站 http://www.tscss.org/主页,点击左下侧的“中国知网”链接,进入该网主页后,先进行注册,用注册名登录后再在页面上点击“我的 CNKI”进行充值,按照提示将所持卡卡号、密码输入,即可将本卡面值充入该账号。使用已充值的用户名、密码登陆,即可实现 CNKI 系列数据库的全文检索,并下载所需文章。