投《术语标准化与信息技术》，2008 年，第 2 期，总第 50 期

香港“双语法例资料系统”法律术语的统计分析

那日松[1]　　揭春雨[2]　　冯志伟[3]

摘要：本文使用计算机对于香港汉英双语法例资料系统的文本语料和法律词汇进行了用字和用词的统计分析，并且还对文本语料中的标点符号进行了统计分析，指出了香港法律术语中也存在着“术语形成的经济律”，并且遵守“Zipf 定律”，同时也指出了香港法律术语在结构上的某些特点。

关键词：法律术语，统计分析，语料，术语形成经济律，Zipf 定律，标点符号

Statistical Analysis of legal terms in Hong Kong Bilingual Laws Information System (BLIS)

Narsu,　Kit Chunyu,　Feng Zhiwei

Abstract：In this paper, the authors analyze the characters, words, terminology and punctuation in Bilingual Laws Information System (BLIS) with statistical approach. The authors discover that the economic law of term formation and Zipf’s law also exisist in BLIS, and they point out some features in punctuation usage in BLIS.

Key words: legal terminology, statistical analysis, corpus, economic law of term formation, Zipf’s law， punctuation

双语法例资料系统 (Bilingual Laws Information System，简称 BLIS)是中华人民共和国香港特别行政区政府建立的一个关于所有现时实施的香港法律的主体条例及附属法例的中文文本和英文文本的资料系统，可以通过网络进行检索。其主页如图 1 所示：

[1] 中国传媒大学
[2] 香港城市大学中文翻译和语言学系
[3] 教育部语言文字应用研究所

图 1　香港双语法例资料系统主页

BLIS 包括如下内容：

1. 香港成文法的中文及英文文本
   - 所有现实实施的香港主体条例及附属法规；
   - 所有香港主体条例及附属法规（包括已经废除的法规）追溯到 1997 年 6 月 30 日为止的过去版本。
2. 宪法类文件、全国性法律及其他有关文件
   - 中华人民共和国宪法、香港特别行政区基本法、全国人民代表大会的有关决定、全国人民代表大会常务委员会的有关决定及解释，以及中英联合声明；
   - 在香港特别行政区实施的全国性法律；
   - 香港特别行政区立法会议事规则。
3. 香港法例所用的词汇用语
   - 英汉法律词汇；
   - 汉英法律词汇。
4. 条例主题索引
   - 条例中文主题索引；
   - 条例英文主题索引。

可以看出，BLIS 是研究香港法律术语的一个非常宝贵的语言资源，我们对于 BLIS 进行了初步的分析。在本文中，我们首先对于 BLIS 资料库中 1000 多万字（包括 21 万个句子）的中文文本语料进行用字、用词的分析，然后再进一步对于 BLIS 资料库中汉英对应香港法律词汇表中的 1 万多条术语进行用字、用词的分析，最后对于标点符号进行分析。

1. BLIS 法律文本语料中的用字和用词统计分析

1.1. BLIS 法律文本语料中用字的统计分析

BLIS 法律文本语料的总字符数量很大，将近 1000 万个字符，其中包含 760 多万个汉字字符和 200 多万个标点符号、非汉字字符和数字。分布情况如表 1 所示：

	汉字字符	其他符号（数字，非汉字字符，标点符号和特殊符号）
语料大小	760 多万（3294 种汉字）	200 多万（165 种其他符号）

表 1　BLIS 法律文本语料中的字符数量

其中，出现频次超过 10 万的字和符号有 11 个，分别是：

的：342426；
）：270113；
（：270112；
，：230335；
。：181581；
或：173946；
1：141918；
第：127095；
该：115523；
人：109432；
条：105180。

只考虑汉字的频次：占 760 多万汉字的 1%以上的汉字有 7 个，分别是：“的，或，第，该，人，条，任”。它们是 BLIS 法律文本中频次最高的汉字，这少量频次最高的汉字对于整个 BLIS 文本的构成具有决定性的作用，很多单词都是由这些少量的高频次汉字构成的。这种使用最小数量的汉字表达最多的文本信息的现象，反映了人类行为的“最小用力原则”（principle at least effort）。在术语学理论研究中，我们应当注意这个重要的原则。

1.2. BLIS 法律文本语料中用词的统计分析

汉语的书面文本是由连续的汉字流组成的，单词与单词之间的界限被淹没在前后相续的汉字流中了，因此，我们在进行用词统计的时候，首先必须“分词”（word segmentation）。我们使用中国科学院计算所的汉语词法分析系统 ICTCLAS 作为软件工具来进行 BLIS 文本的自动分词。自动分词之后，我们获得了 16190 个词语组合，可以分为两类：一类是特殊组合术语，一类是完全由汉字组成的术语。

特殊组合术语可以分为三类：

1. 标点符号和汉字的组合：例如，“隔舱”式。
2. 英文字符和汉字的组合：例如，Bowling 型，Deighton 型。
3. 英文字符、数字和汉字的组合：例如，HM527 型，IIC 级。

完全由汉字组成的词语，可以按照汉字数目的多少，分为单字词，二字组合，三字组合，四字组合，…，等。统计分析结果如下：

单字词排在词汇频次统计的前列，例如：“的，或，该，在，人，须，而，及，所，年，可”。前三个字的频次都在 10 万以上。

频次比较高的二字组合有：“任何，根据，条例，作出，规定，其它，法律，有关，修

订，注册，命令，情况，公司，不得，指明，公告，申请，提出，发出，进行，使用，程序，通知，香港，条文，日期，船舶，法院”等等。这些都是频次在 8 千以上的二字词汇，它们大部分是双音节的一般性单词，很少有词组型术语。

三字组合有：“委员会，任何人，证明书，通知书，申请人，建筑物，许可证，管理局，退休金，代理人”等等。它们的频次都在 1900 次以上。

四字组合有（括号中的数字是频次）：“切实可行（1667），另有所指（949），有限公司（327），法定人数（231），无线电话（217），无线电报（157），上层建筑（99），任何法院（94），明文规定（93）”等等。

五字组合有（括号中的数字是频次）：“任何委员会（139），维多利亚港（29），印度尼西亚（16）”等等。

六字和六字以上组合（括号中的数字是频次）：“中华人民共和国（113），任何附属公司（26），香港中文大学（33），香港国际机场（25），香港房屋委员会（12），欧洲经济委员会（11），中华电力有限公司（9），香港上海汇丰银行有限公司（8），香港联合交易所（8），香港加德士有限公司（8）,欧洲经济共同体（8），香港电话有限公司（6），英国工程标准协会（5），巴布达利比利亚利比亚伯利兹（1）”等等。

这些数据说明，双字组合的词语所占比率很高，而三字和四字，以及多字组合词语的使用次数越来越少。六字和六字以上组合词语基本上都是各种组织机构名称，以及专有名词。由于我们使用的分词工具 ICTCLAS 是面向一般性文本的，因此，切分出来的组合除了术语之外，还有一些一般性的词语，因而使得双字组合的词语占了大多数，而且它们很多都是双音节的一般性单词，这些双音节的一般性单词与双音节的单词型术语用我们的分词工具是难以区分开来的。由于分词工具的这种局限性，我们的统计结果只是反映了汉语书面语中双音节词占大多数的语言事实，还没有反映出法律术语在音节分布上的特殊情况。

2. BLIS 汉英法律词汇表中术语的用字和用词统计分析

为了找出法律术语在音节分布上的特殊情况，我们以汉英法律词汇表中的术语作为统计对象，由于我们的统计对象中不再包含一般性的词语，我们的统计结果就能反映出法律术语的特点。下面，介绍我们对于 BLIS 汉英法律词汇表中的术语的统计分析结果。

2.1. BLIS 汉英法律词汇表中术语的用字统计分析

统计分析是在 9652 条汉英法律词汇表里的术语进行的（如果多义术语的不同义项分别算为不同的术语，那么，这个法律词汇表里的术语共有 1,0477 条），其中包含汉字字符和其他符号（标点符号、数字和非汉字字符）的术语有 1336 条，全部由汉字字符组成的术语有 8216 条。在这些术语中，频次比较高的几个汉字是：

的：1188，
法：1161，
人：971，
权：909，
证：635，
行：528，
有：522，
律：510

法律词汇表中术语的高频字使用情况和 760 万文本语料中的高频字有很多交集，这些字不管是在法律词汇表里还是法律文本里都是高频字。如表 2 所示。

汉字	的	法	人	权
语料中	342426（4.5%）	53943（0.7%）	109432（1.4%）	29139（0.38%）
术语表中	1188（2.4%）	1161（2.4%）	971（2.0%）	909（1.9%）
汉字	证	行	有	律
语料中	28455（0.38%）	47798（0.62%）	85184（1.1%）	26258（0.35%）
术语表中	635（1.3%）	528（1.1%）	522（1.1%）	510（1.0%）

表 2.　文本语料与术语表中的用字比较

从分析结果可以看出，不论在文本语料还是在法律词汇表中，高频字构成术语的比例比低频字构成术语的比例大得多。频次高的汉字占了整个文本语料以及法律词汇表的大部分，只出现过一次或者几次的低频汉字数量不小，但覆盖面不大。

根据我们的统计，在 760 多万个汉字的 BLIS 文本语料中一共只使用了 3294 个汉字，这意味着，只要不到半数的国家标准汉字（GB2312-80 国家标准中共有 6763 个汉字）就可以组合成 BLIS 全部的法律术语。可见法律术语中的汉字也遵从由少量的高频汉字组成大量的法律术语的“最小用力原则”。有的学者指出，法律术语具有“简练性” ，他们认为，法律术语使用最少量的汉字来表达最大的信息量，用最简练、明晰的语言材料来传递最充分完备的信息[4]。其实，如果把他们说的这种“简练性”用数学方法从理论的角度加以抽象，就可以提升为反映语言现象客观规律的定律，在数理语言学中叫做“Zipf 定律”（Zipf’s law）[5]。对这个问题有兴趣的读者，可以进一步参看冯志伟的《数理语言学》，这里不再赘述。

2.2.　BLIS 汉英法律词汇表中的用词统计分析

从结构上看，法律术语主要以“偏正”、“并列”、“支配”，“动宾”等结构方式出现，“主谓”结构的词组比较少。这一点，在 BLIS 文本语料和在法律词汇表中的表现都是一样的。法律术语这种在结构上的特点与医学术语是很不一样的，医学术语中的主谓结构很多，如“头痛，肺结核、胆结石”，但是在法律术语中，这种主谓结构的术语很少见。

BLIS 法律词汇中的术语，在词性方面主要以名词和名词词组（例如，“法律，法官，法庭”等）为主，也有一些动词或动词词组（例如，“传召、辩护、送达、聆讯”等）。在计算机科学的术语中，基本上是名词或名词词组（例如，“软件，硬件，程序，算法”等），冯志伟在《现代术语学》中曾经指出，计算机术语也可以是动词[6]（例如，“打印、走纸、导出、译码、消零、置零、回溯”等）。但由于传统的术语学是基于概念的，大多数学者们只研究反映概念的名词或名词词组，有的学者干脆把“术语”叫做“名词”，因此，冯志伟的这种意见遭到一些学者的反对。从 BLIS 法律词汇中的术语看来，法律术语中有不少的动词术语，冯志伟的意见似乎是有道理的，我们建议不要轻易地否定他的意见，进一步根据大量的语言事实来研究这个问题。

BLIS 汉英法律词汇表中收录的全部都是术语，从这些术语的长度考察，以四字组合到六字组合的术语居多，而二字组合的术语明显地比文本语料库中的二字组合术语减少了。具体地说，二字术语只有 1004 条，而四字组合术语有 3289 条，五字组合术语有 1481 条，六字组合术语有 1199 条，它们都远远地超出了二字术语的出现次数。三字组合的术语数量较少，只有 704 条。

由于汉语书面语中的双音节词占大多数，所以，二字组合术语主要是单词型术语，三字组合术语也主要是单词型术语，而四字组合术语、五字组合术语、六字组合术语主要是词组型术

[4] 安秀萍，法律文书语汇初探，山西省政法管理干部学院学报，2006.12，19(4)，21-23 页

[5] 冯志伟，数理语言学，上海知识出版社，1985 年，第 151 页。

[6] 冯志伟，《现代术语学引论》，语文出版社，1997 年，第 116 页。

语，因此，我们可以说，在 BLIS 的法律词汇表中的术语，词组型术语占了大多数，这种情况与 BLIS 文本语料中的单词长度的分布是大不一样的（双音节词语占大多数），反映了法律术语结构在长度方面的特点，而且，也符合“术语形成经济律”。

冯志伟在《现代术语学》提出了“术语形成经济律”，他指出：术语系统的经济指数、单词的术语构成频度以及术语的平均长度之间，存在着相互依存和相互制约的关系；在一个术语系统中，提高术语系统经济指数的最好方法，是在尽量不过大改变术语平均长度的前提下，增加单词的术语构成频度，尽量使用原有的单词型术语来构成各种词组型术语；这样，在一个术语系统中，词组型术语的数量必定会超过单词型术语的数量[7]。

BLIS 法律词汇表中的术语长度的统计结果说明，法律术语中的词组型术语占了大多数，大部分术语的长度分布在四字到六字组合的术语之间，它们基本上是词组型术语，而不是单词型术语。这种情况，进一步证实了冯志伟提出的“术语形成经济律”确实是术语学中的一个重要规律，这是我国学者对于现代术语学基础理论的深刻思考和重要贡献，对于术语工作具有指导意义。“术语形成经济律”是冯志伟在 20 世纪 80 年代提出的，直到今天，国外术语学文献还没有见过任何与此类似研究的报道，这是我国学者独创性的理论研究成果。

下面我们给出 BLIS 法律词汇表中的术语长度和出现次数的统计结果：

词汇长度	出现次数	词汇长度	出现次数	词汇长度	出现次数
4	3289	5	1481	6	1199
2	1004	7	802	3	704
8	391	9	278	10	171
11	115	12	86	13	40
14	22	1	19	16	15
15	13	19	8	17	5
22	3	18	2	27	2
29	1	21	1	25	1

表 3　法律词汇表中术语长度的出现次数

3.　法律术语和标点符号的关系

在 BLIS 语料中，有相当的一部分术语是以术语跟随术语解释的形式存在的。我们抽取了这一部分术语。已经校对过，共 466 条。例如：

1. “经修订的条例”指经 1998 年第 26 号修订的第 1 章。
2. “个案 1”指债务人是未获解除破产的破产人的个案。
3. “个案 2”指债务人并非未获解除破产的破产人的个案。
4. “条例”(the Ordinance)指《公司条例》(第 32 章)。
5. “《1978 年国家豁免权法令》”乃“State Immunity Act1978”之译名。
6. “《基本法》”(Basic Law)指《中华人民共和国香港特别行政区基本法》。
7. “成人”、“成年人”(adult)*指年满 18 岁的人。
8. “月”(month)指公历月。

带双引号的部分就是 BLIS 的术语。从这些双引号术语的结构特点来看，在左右的双引号之间，是汉字或数字。由于这些汉字或数字有了标点符号的边界修饰使得其术语的特点更加直

[7] 冯志伟，《现代术语学引论》，语文出版社，1997 年，第 128 页。

观和明显了。

在 BLIS 的文本语料中，有些词语只有加了特殊的标点符号标记之后才能被认定为术语，例如，很多术语都使用双引号特别地表示出来，因此，我们有必要研究术语与标点符号的关系。下面介绍我们研究法律术语和标点符号间的关系所做的部分试验。

如果只考虑 BLIS 文本中标点符号的出现频次，按从高到低的顺序排列如下：

左右括号>逗号>句号>分号>顿号>双引号>书名号

其他标点符号用得比较少。

封鹏程在他的硕士论文“现代汉语法律语料库建立及其词汇计量研究”[8]中，对内地法律文本中标点符号的统计得出的结论是：在内地法律文本中，逗号和句号的出现频次远远超过左右括号出现的频次。但是在香港法律文献 BLIS 中，左右括号的使用频次却排在逗号和句号的前面，其频次超出了逗号和句号，这是 BLIS 标点符号使用的重要特点。这意味着，尽管同样是法律文献，内地法律文献和香港法律文献在标点符号的使用上存在着明显的差异：在香港法律文献中，左右括号的使用频度非常高，使得香港法律文献具有一种迥然不同的标点符号使用风格。

此外，根据我们的统计，在香港法律文献中，双引号和书名号使用频度也比内地法律文献为高。

在我们自动抽取法律术语的时候，如果考虑这些标点符号的特殊的标示作用，把它们作为有关的术语的形式标志，就有助于计算机自动地从 BLIS 文本语料中抽取到术语。根据这样的特点抽取出来的术语可以形成自动抽取 BLIS 法律术语的基础术语库。从这个基础术语库出发，以它们作为“种子”，再使用其他的方法，我们就可以实现 BLIS 法律术语的自动抽取，这是我们目前正在进行的一项研究工作。

参考文献

1. 冯志伟，现代术语学引论，语文出版社，1997 年。
2. Heribert Picht, Jennifer Draskau, Terminology: An Introduction, University of Surrey England, 1985 .

[8] 封鹏程，现代汉语法律语料库的建立及其词汇计量研究，南京师范大学，硕士论文，2005。