基于银行语料库文本的简要分析

* 1. 类符/形符比 (TTR)

类符 (type) 指的是语料库中不同单词的数量, 而形符 (token) 指的是语料库的总词数。类符形符比 (TTR) 就是类符和形符之间的比率。单纯的类符数和形符数不能反映文本的本质特征, 但两者的比率在一定程度上反映了文本的某种本质特征, 即用词的变化性 (杨惠中, 2002) 。比值越高, 说明语料使用的词汇越灵活丰富。但是, TTR无法科学反映大小不同文本的用词丰富程度。标准化类符形符比(STTR) 则是按一定长度 (通常是1 000词) 分批计算文本的类符形符比, 然后算出平均值。因此, 对于字数超过1 000的文本, 学界一般使用标准化TTR来衡量其词汇使用的变化情况, 以便减少虚词词频过高所导致的误差。

|  |  |  |
| --- | --- | --- |
|  | BC |  |
| 类符数 | 483 |  |
| 形符数 | 1023 |  |
| 标准化类符形符比（STTR） | 47.3 |  |

* 1. 平均词长

平均词长指的是文本所有单词的平均长度。平均词长能够反映文本词汇的复杂程度。平均词长越长, 说明文本中使用的长词越多。普通文本的平均词长大约为四个字母, 低于四个说明文本用词相对简单, 远高于四个则说明文本用词较为复杂。根据wordsmith的统计，收集到的银行文本的平均词长为5，文本用词难度适中，较为接近普通文本。

Word Smith软件能够快速统计出各长度词出现的次数。

|  |  |  |
| --- | --- | --- |
| 词长 | 词频/次 | 百分比 |
| 1-letter words | 25 |  |
| 2-letter words | 181 |  |
| 3-letter words | 232 |  |
| 4-letter words | 174 |  |
| 5-letter words | 117 |  |
| 6-letter words | 81 |  |
| 7-letter words | 84 |  |
| 8-letterwords | 46 |  |
| 9-letter words | 35 |  |
| 10-letter words | 22 |  |
| 11-letter words | 18 |  |

上表统计数据显示, 三个字母的单词在语料中出现的频率最高，近80%的词汇都是由2~6个字母组成的简单词汇构成。这说明语料用词有口语化倾向，符合贷款说明文体的文体特征。

* 1. 词汇密度
  2. 高频词分布

2.1 平均句长

平均句长就是文本中所有句子的平均长度。就整个语料库而言, 句子的长短在一定程度上反映了句子的复杂程度。一般而言, 平均句长越长, 文本的句法结构就越复杂, 句子的理解难度越高。Butler曾按长度把句子分为三类:短句 (1~9个词) 、中等长度句 (10~25个词) 和长句 (25个词以上) 。该银行语料的句子信息统计结果如下：

|  |  |
| --- | --- |
|  |  |
| 句子个数 | 21 |
| 平均句长 | 49 |
| 句子标准差 | 42.47 |

3. 语篇层面

3.1 文本正式化程度

3.1.1 名词化现象

名词化是指通过添加词缀或转化等方式把其他某个词类转化为名词的语法过程。名词化现象是英语语言中一种极为常见的现象, 同时也是英语正式语体的一个显著特征。英语名词化的形式很多, 如后缀派生法和零位派生法。本研究仅就几个常见的由后缀派生法生成的名词化形式进行统计分析, 包括-ment;-tion;-ity;-ness;-sion。本文利用Wordsmith工具的Concord功能上述五个后缀进行检索, 并人工剔除不符合要求的检索结果。检索结果如下：

|  |  |
| --- | --- |
| 后缀 |  |
| -tion(s) |  |
| -ment(s) |  |
| -ity(ities) |  |
| -ness |  |
| -sion(s) |  |
| 总计 |  |
| 占总词数百分比 |  |

3.1.2 被动语态的使用

汉英两种语言在主动和被动两种语态的使用方面存在较大的差异。在汉语中, “被”字式往往用来表达对主语而言不如意甚至不幸的事, 故曾被称为“不幸语态”。 汉语当中往往避免使用“被”字, 在行文中多选择从主动角度表达其被动含义的词语。相反, 被动语态在英语中使用频率很高, 尤其是在比较正式的文体当中, 比如科技文体、政府公文和新闻报道。人们使用被动语态往往是为了表达客观性、正规性、中立性和礼貌等。被动语态的使用频率可以从侧面反映文本的正式程度。一般而言, 被动语态使用越频繁, 代表文体越正式。

3.2 衔接和连贯

语篇的衔接和连贯是语篇研究的两个核心概念。衔接体现语篇的表层结构关系, 是语篇的有形网络。语义连贯体现语篇深层内容上的连接, 是语篇的无形网络。由于篇幅有限, 本文仅从人称指称和连词两个方面考察语料的衔接手段使用情况。

3.1.3 连词使用情况

本文利用WordSmith软件中的WordList功能, 对语料中出现频数最高的十个连词进行了统计, 结果如下：

|  |  |  |
| --- | --- | --- |
| 连词 | 词频 | 百分比 |
| And | 42 | 4.10 |
| When |  |  |
| But |  |  |
| If |  |  |
| Or | 2 | 0.2 |
| Though |  |  |
|  |  |  |

3.1.4 人称代词主语的使用

本研究利用Wordlist的Concord功能对语料中出现的人称代词主语进行了检索, 并人工剔除you和it的宾格形式, 结果如下:

|  |  |  |
| --- | --- | --- |
| 人称代词主语 | 词频/次 | 占总词数百分比 |
| His | 9 | 0.88 |
| I | 0 | 0.888 |
| You | 24 | 2.34 |
| It | 16 | 1.56 |
| Them | 6 | 0.59 |
| We | 4 | 0.39 |

3.3 可读性分析