**基于语料库的银行贷款网站文本的简要分析**

为了进行案例分析和基于语料库的词汇表达方式研究，我们从国外八大银行官方网站私人贷款介绍页面的表述文本为研究对象，进行统一的文本提取，汇总，处理，自建语料库，在充分阅读和了论文中对于各种语体分析和语言统计分析对象和方法的基础之上，使用word smith，sketch engine，readability analyzer等语料库分析软件，从词汇，句子，篇章三个层面系统分析了国外贷款网站文本使用上的特点以及分析了值得我们本地化改写可借鉴的地方，也使用了术语提取工具和语料库索引工具提取了来自于这些文本的专业术语表，关键词表和高频词表留作下一步本地化写作素材库的制备和写作过程的支援。具体分析如下所示：

1. **词汇层面**
   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1000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1000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。使用wordsmith计算得到的类符，形符数量及其比值如下：

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

从上表数据可以看出，该银行贷款汇总语料的标准化类符形符比为47.3%，是一个比较高的比值，说明贷款语料的词汇变化较为丰富, 可读性也较高。

* 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4.88，文本用词难度适中，较为接近普通文本。

Word Smith软件还能够快速统计出各长度词出现的次数。

|  |  |
| --- | --- |
| 词长 | 词频/次 |
| 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**

词汇密度就是实词在文本总词数中所占的百分比。实词的意义更为稳定, 能传达更多信息, 而虚词一般起语法和文本的连接作用。词汇密度的高低能够说明文本信息量的大小和难易程度。一个文本的词汇密度越高, 说明实词与总词数之间的比值越大, 文本所承载的信息量也就越大, 文本的阅读难度也就越高。计算词汇密度的方式。相较而言, 应用更为广泛的是Ure(1971)计算公式, 即词汇密度=实词数/总词数×100%。参照Ure(1971)计算公式，使用sketch engine自建语料库分词性检索和统计如下:

|  |  |
| --- | --- |
| 词性 | 出现频率 |
| 名词 | 1305 |
| 动词 | 715 |
| 形容词 | 262 |
| 副词 | 122 |
| 实词数 | 2404 |
| 总词数 | 4662 |
| 词汇密度 | 51.5% |

Laviosa (1998) 的研究发现英语源语文本的词汇密度 (54.95%)。而上表数据显示，银行语料的词汇密度仅为51.5%，低于正常英语源文本的词汇密度，说明银行贷款语料追求简单易懂，尽量降低文本的信息负载量，以减少客户的阅读负担和阅读难度。

**2. 句子层面**

**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。(杨惠中，2002：135)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 (1985:121)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 | 541 |
| 平均句长 | 11.71 |

上表数据显示，语料的平均句长为11.71，较多使用中短句子，避免长句, 有助于读者的理解。

**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sketch engine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) | 52 |
| -ment(s) | 94 |
| -ity(ities) | 24 |
| -ness | 2 |
| -sion(s) | 3 |
| 总计 | 175 |
| 占总词数百分比 | 3.7% |

由上表数据显示，国外银行贷款介绍呈现明显的名词化特征，文体的正式化程度比较高，我们在进行本地化写作时也要尽量避免动词的大量使用，尽量选用名词，体现文本的专业化和正式化程度。

**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。

通过在Datayze.com网页的免费统计发现，语料中出现的被动语态频数为49次，被动语态的使用也是较频繁的。

**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。

**3.2.1 连词使用情况**

英语不论是词语、短语还是分句之间的衔接都需要通过连词来实现。因此, 在进行本地化改写时, 通常需要增加and、though、while、if、therefore等连词以明示原文隐含的逻辑关系。

本文利用WordSmith软件中的WordList功能, 也对从各网站搜集来的国外银行贷款汇总建立的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 | 170 | 3.5 |
| But | 4 | 0.08 |
| If | 58 | 1.2 |
| Or | 78 | 1.6 |
| So | 10 | 0.2 |
| As | 55 | 1.1 |
| Either | 3 | 0.06 |
| Than | 11 | 0.2 |
| Both | 3 | 0.06 |
| Because | 3 | 0.06 |
| 总计 | 395 | 8.06 |

从上表数据不难看出, 使用频率最高的前十个连词总频数都占总词数的8.06%，这说明原银行贷款语料比较注重通过添加连词的方式将隐含的逻辑关系明示出来, 语篇连贯通顺。

**3.2.2 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 |
| He | 9 | 0.02 |
| I | 0 | 0 |
| You | 217 | 4.5 |
| It | 28 | 0.5 |
| They | 10 | 0.2 |
| We | 63 | 1.3 |

从统计数据可以看出，语料中使用最多的两个代词是you 和 we， 说明银行贷款语料会用“you”的第二人称站在客户角度为客户介绍申请条件，当介绍银行本身对客户的承诺时会用“we“第一人称拉近与客户的距离。

**3.3 可读性分析**

语篇可读性指的是语篇的阅读难易程度，可在宏观层面上检验语料是否呈现简化趋势。常用的语篇可读性的计算公式有三种:福来士易读指数 (Flesch-Kincaid Reading Ease) 、福来士难度级数 (Flesch-Kincaid Grade Level) 、迷雾指数 (Gunning Fog Scale Level) 。福来士易读指数按百分制评定语篇的阅读难度, 分值越高说明语篇越容易理解。65是一般难度英语的分值, 说明文本的平均句长为15到20个单词, 平均词长为两个音节。得分在60-70范围内对应于8年级/ 9年级英语水平。50到60之间的分数对应于10/12年级。30以下是大学毕业生。福来士难度级数的测量基于美国学校年级等级, 如7.0表示七年级学生能看懂的语篇, 一般7.8到8.0之间属于正常水平。迷雾指数常用来测算读者需要具备多少年的正规教育才能轻松读懂一篇文章。迷雾指数为6代表文章阅读难度不高, 读者只需接受六年教育便可轻松读懂。指数越高代表阅读难度越高。5是可读的，10是适中的，15是比较困难的，20是非常困难的。将语料上传到Datayze.com进行相关的可读性检测结果如下：

|  |  |
| --- | --- |
| 语篇可读性计算公式 |  |
| Flesch-Kincaid Reading Ease | 59.27 |
| Flesch-Kincaid Grade Level | 7.9 |
| Gunning Fog Scale Level | 11.07 |

上表数据显示, 三种可读性计算公式都得出一致的结论，语料的可读性都是比较适应于中学或大学生阅读的正常水平，这说明我们在进行本地化改写时，要注意用词既不能过于简单也不能使用一些生涩偏僻的词汇，风格要以简洁明了为主。

**参考文献：**

[1] Laviosa, Sara.1998.Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose [J]. Target (1): 23-48

[2] Ure, J.N.1971.Lexical Density and Register Differentiation[G] G.Perren&J.L.M.Trim Applications of Linguistics Cambridge: Cambridge University Press.

[3] Butler, Christopher.1985.Statistics in Linguistics[M].Oxford:Basil Blackwell.

[4] 黄立波.2018.语料库译者风格研究反思.[J].外语教学（1）：77-81

[5] 王晋军.2003.名词化在语篇类型中的体现[J].外语学刊（2）：74-78.

[6] 杨惠中.2002.语料库语言学导论[M].上海.上海外语教育出版社.