基于语料库的文学作品分析方法初探

华南师范大学南海校区英语系 肖普勤* 黄凤枝**

摘 要：语料库检索软件检索能力强大，可用在文学文本的分析上，用语料库方法分析文学作品能揭示文本的主题是如何通过故事情节、人物刻画、修辞手法等来表达的。本文探讨了基于语料库的文学作品分析方法。该方法不仅可以用于文学批评领域的文本分析，还可以用于涉及文本分析的语言课堂教学。

关键词：语料库 检索 关键词 文学作品 分析方法

一、引 言

20 世纪 90 年代以来语料库语言学快速发展，语料库分析方法在文学批评领域的运用日益增多，通过检索软件获得的语料被认为可以为批评家的直觉提供数据验证（林丽云，2004）。语料库检索软件检索能力强大，可用在文学文本的分析上，以此“把隐含的结构显现出来，同时激发人的想象力，并能检验文本对读者的感染力”（杨建枚，2002）。国内已有研究者运用语料库方法来分析文学作品，例如，郭放所做的“《快乐王子》的语料库检索分析”、杨建枚所做的“《警察与赞美诗》的语料库检索分析”和张厚振所做的“基于语料库的海明威作品《一个干净、明亮的地方》分析”（参见参考书目）。用语料库方法分析文学作品能揭示文本的主题是如何通过故事情节、人物刻画和修辞手法等来表达的，但是语料库语言学对于相当多的语言学研究者还很陌生，更不用说如何将语料库分析方法用于文本分析或文学作品分析了。因此，有必要和对语料库语言学感兴趣的研究人员、教师和学生一起探讨基于语料库的文学作品分析方法。本文从文本总体统计特征和分析、主题和情节检索与分析、对人物刻画的检索与分析、对修辞手法的检索与分析四个方面简要说明基于语料库的文学作品检索步骤和分析方法。

二、文本总体统计特征和分析

基于语料库的语言研究一般采取定性与定量相结合的研究方法，要进行定量研究就要涉及文本的检索和数据的统计。语料库分析软件一般对文件进行操作。如果每一个文本是一个独立的文件，那么可以统计出每一文本的总体统计特征及整个语料库的总体统计特征。如果文本是分类存储的，那么可以统计出每一类的总体统计特征及整个语料库的总体统计特征。如果语料库是一个文件，那么只能统计出整个语料库的总体统计特征。

主要的统计特征有：文件的字节数（bytes）、形符数（tokens，指文本一共有多少个词）、类符数（types，指文本一共有多少个不同的词形）、类符形符比（type/token ratio）、标准化类符形符比（standard type/token ratio）、平均词长（average word length）、句子数（sentences）、平均句长（sentence length）、句长标准差（standard deviation of sentence length）、段落数（paragraphs）、平均段落长（paragraph length）、段落长标准差（standard deviation of paragraph length）等等（杨惠中，2002）。

从文本的字节数、形符数和句子数可以推断文本的篇幅长度。单纯的形符数和类符数不能反映文本的

* 肖普勤（1976-），华南师范大学外文学院硕士研究生，华南师范大学南海校区英语系助教；研究方向：翻译，语料库语言学；通讯地址：华南师范大学南海校区英语系，邮编：528225；E-mail: seanxpq@163.com, seanxpq@126.com.
** 黄凤枝（1977-），华南师范大学外文学院硕士研究生，华南师范大学南海校区英语系助教；研究方向：翻译，应用语言学。

本质特征，但两者的比率却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。一般说来，类符形符比越高，用词变化性越高。但英语的词汇是有限的，如果文本不断扩大，形符数将随之扩大，然而类符数的增加却不能保持同步，所以当文本容量达到一定程度时，类符数的增加将越来越小，两者的比率无法反映用词的变化性。因此我们需要采用标准化类符形符比来反映用词的变化性，其计算方法是按一定的长度分批计算文本的类符形符比，然后求出它们的平均值（杨惠中，2002）。利用平均句长和句长标准差可以判断文本的句子是否比一些简易文本句子要长。同理，我们还可以用它来比较段落长。根据平均词长，可以计算低于该词长的类符数在总类符数中所占比例，以判定该文本的词汇难度。具体实例可参见杨建枚的“《警察与赞美诗》的语料库检索分析”（杨建枚，2002）。

三、主题和情节检索与分析

对作品用语料库软件（如 AntConc 3.01）生成词表（wordlist），可以得到文本中出现频率最高的词的词频（以词的频率多少排列）。对这些词的初步分析可以让我们了解有关文本内容的信息，但无法确定哪些信息重要，哪些信息次要。我们需要选择关键词（key words）以确定关键信息。一个词是否是某一文本或文类（genre）的关键词，不仅取决于该词在该文本或文类中的出现情况，还取决于该词在与之相对比的参照语料库中的出现情况。假如定冠词 the 在某一长度为 1000 词的文本中出现了 50 次，其出现频率达到了 5%，但不能说 the 是该文本的关键词，这是因为 the 在任何文本中的频率都很高，不是惟独在这一文本中出现频率高，它在参照语料库中的频率可能还不止 5%，单就其在该文本或文类中出现的频率来决定它是否关键词显然是不合适的。（杨惠中，2002）因此，我们可以选取某一参照语料库（长于被检索文本），按照关键值（keyness value）生成关键词词表或主题词词表（keyword list）。通过观察关键词词表中排在较前的关键词，可以得到该文本的最关键信息（关键值越高说明该信息越重要），如故事中的主题、主要人物、时间、地点、背景、关键描述信息等。

为了了解文本（如小说）的大致情节，我们可以利用语料库软件的语境共现（concordance）功能，输入关键词（如主要人物）进行带语境的关键词（指搜索词）（KWIC）检索。通过分析并阅读搜索词两边的语境（或上下文），就可以得出作品的主要情节了。具体实例可参见郭放的“《快乐王子》的语料库检索分析”（郭放，2004）。

Wordsmith Tools 在提取主题词和了解作品的概况方面有独到的优势。首先，它可以与参照语料库对比生成一个按照关键值排列的主题词表（与上述的 AntConc 3.01 操作过程相似）。对主题词的分析过程与上述相同。其次，它有独特的词图（plot）功能。词图统计是根据主题词表，计算出各个主题词在语篇中的位置分布，其意义主要在于对某一连续文本的词语分布进行统计和计算。尽管其他软件（如 AntConc 3.01）也有词图功能，但只能统计单个词的词图。单个词图只有同其他词图放在一起进行比较，才显示出真正意义。对 Wordsmith Tools 产生的词图进行观察就可以直观地、清楚地看到故事情节的开始、发展、高潮、结局等各环节。主题词在词图中所体现出来的密集与稀疏真实地反映了各条线索的发展，因而对于了解文本的情节有明显的优势。具体实例可参见张厚振的“基于语料库的海明威作品《一个干净、明亮的地方》分析”（张厚振，2004）。

四、对人物刻画的检索与分析

作品中的主要和次要人物一般会出现在主题词表（或关键词表）的较前位置。欲认识各人物形象，可依次输入人物关键词（如表示姓名的名词和其人称代词、形容词性物主代词等），检索全文，定能从检索项的共现语境中查到相关的名词、形容词、动词、副词以及短语等。这些词和短语就是用来修饰和限制检索项的。

将所收集到的词和短语进行分类、归纳和分析（如积极与消极、正面与负面），就能够描述人物的外貌、活动、性格、心理活动等方面了。具体实例可参见郭放的“《快乐王子》的语料库检索分析”（郭放，2004）。

五、对修辞手法的检索与分析

在对文本有了初步的认识后，再进行修辞手法的检索与分析是比较合适的。语料库软件并没有自动识别、检索修辞手法的功能。因此，我们可以根据对修辞手法各种特征的了解和对试读语料的主观印象及从语料中发现的个例，提出可能的检索项，充分运用语料库软件的检索、计算功能找到各种修辞实例。具体实例可参见郭放的“《快乐王子》的语料库检索分析”（郭放，2004）。

首先，最容易找到的修辞手法应该是明喻（simile）。它利用不同事物之间的相似点，借助比喻词（如 like, as）起连接作用，清楚地说明甲事物在某方面象乙事物。有几种类型：(1) like 型；(2) as 型；(3) 虚拟句型（最常见的是 as if 或 as though，might have done/been）；(4) what 型（常用句式：A is to B what X is to Y; What X is to Y, A is to B）；(5) than 型；(6) and 型。（李冀宏，2000）因此，检索以上各词和词组，并分析结果即可。

第二，隐喻直接将甲事物当作乙事物来描写，无须借助比喻词。有几种类型：(1) 名词型（最常见的句式是“甲是乙”，喻体一般体现在句子的标语部分）；(2) 动词型；(3) 形容词型；(4) -of-短语型（李冀宏，2000）。对于名词型的隐喻，可以检索 be 的各种形式；对于-of-短语型隐喻，可以检索 of；但对于动词型和形容词型隐喻，就只能通过阅读语料凭主观判定了。

第三，排比的构成可体现于各个语言层次，如单词、短语、从句、句子等，其中以三项式平行结构最为普遍（李冀宏，2000）。它的平行特点与其标点符号紧密相关，因此，可以搜索词位置为中心，限定其检索的跨距（span），检索“，；。”。

其它的修辞手法可能很难用语料库手段进行发掘，问题的关键在于要检索的搜索项无法自动确定，必须人工确定。但一旦确定搜索项，语料库所提供的方便与快捷是无与伦比的。

六、结 语

基于语料库的文学作品分析方法“虽然在语言特征判断方面无重大突破，但实施起来却快捷、准确，省时省力，而且证据充实，令人信服。”（何安平，2001）该方法不仅可以用于文学批评领域的文本分析，还可以用于涉及文本分析的语言课堂教学。林丽云做了尝试，将语料检索运用于英语精读课的教学中：通过语境共现揭示语篇主题发展和变化，通过词汇复现突现语篇的文体和语言特征；在引导学生学习词汇的同时，从语篇、语言特点、及文章主题思想方面来阅读、欣赏所学文章（林丽云，2004）。这无疑为我们将语料库语言学研究与实际教学理念、方式、手段的革新结合起来提供了优秀的范例。

参考文献：

1. 郭　放. 《快乐王子》的语料库检索分析[J]. 乐山师范学院学报. 2004(6): 66-69.
2. 何安平. 《用语料库研究语言》导读[A]. Jenny Thomas et al. 用语料库研究语言[M]. 北京: 外语教学与研究出版社. 2001.
3. 李冀宏. 英语常用修辞入门[M]. 上海: 世界图书出版公司. 2000.
4. 林丽云. 语料检索在英语精读课教学中的运用[A]. 何安平. 语料库在外语教育中的应用——理论与实践[C]. 广州: 广东高等教育出版社. 2004.
5. 杨惠中. 语料库语言学导论[M]. 上海: 上海外语教育出版社. 2002.
6. 杨建枚. 《警察与赞美诗》的语料库检索分析[J]. 四川外语学院学报. 2002(3): 56-59.
7. 张厚振. 基于语料库的海明威作品《一个干净、明亮的地方》分析[J]. 新乡教育学院学报. 2004(2): 61-63.

（责任编辑：周　化、左燕红、李　锋）