词汇语义分析在文学计量研究中的运用

王 永

（浙江大学外国语学院，杭州 310058）

提 要：在俄罗斯，运用统计方法研究文学作品已有一百多年的历史，不过，较多集中于诗歌格律、作家风格、作者甄别等研究领域。我国的俄罗斯文学计量研究，尚处于起步阶段。本论文尝试将数据统计分析运用于诗歌研究，以俄罗斯国家语料库为数据采集来源，以词汇语义特征为检索入口，通过统计曼德尔施塔姆及赫列勃尼科夫两位诗人文本的人名、地名数据，对诗歌文本的词汇语义特征做统计分析。结果表明，基于词汇语义统计分析的文学文本研究，有助于揭示作品的创作特征及作者的创作主旨。

关键词：词汇语义类别；诗歌；计量研究

中图分类号：I512.072 **文献标识码：A**

1 引言

长期以来，我国的俄语语言学与俄罗斯文学研究各自为政，鲜有“越界”。近年来，跨学科研究的热潮极大促进了语言学与心理学、人类学、脑科学的交叉，打开了文学与地理学、法学、经济学互通的大门。然而，对于文学与语言学及自然科学的交叉研究，俄语界尚未给予较大关注。事实上，俄罗斯语言学家早在一百多年前就尝试将统计方法用于文学作品的语言研究，在语言、文学、自然科学的“三位一体”研究领域产出了丰硕的成果。如莫罗佐夫（Н.А. Морозов）在“语言学光谱”一文中采用计算方法分析了普希金、果戈里、托尔斯泰等作家作品的语言特征；列斯基斯（Г.А. Лесскис）通过对 19 世纪 7 位作家 11 部心理小说的语言特征进行统计分析，发现了作家的一些个体特征；雅尔霍（Б.И. Ярхо）则致力于建造“精密文学大厦”，认为文学语言具有多种特征，而这些特征是可测量的。[1]

不过，统计方法在文学研究中的运用，主要还在于文学语言的形式方面，如元音辅音交替、词长、句长、词类分布等特征。从俄罗斯学者的研究成果可以看出，文学计量研究最擅长的研究问题就是诗歌格律、民间故事的情节、作品作者甄别等领域。然而，对于文学研究而言，挖掘文学形象的特征、作者的思想内涵及其创作主旨等更为重要，因为“个性化永远是文学艺术创作的追求，也是研究者深入探讨的问题所在”（王永 2021：649）。而要发现文学作品的创作个性，应从语言形式层面深入到其语义层面。近年来，我们通过挖掘诗歌作品的语义类别及其词频统计分析，在基于词汇语义统计分析的诗歌研究方面做了一些尝试。语言是文学的载体，通过语言分析，可以将作品层层剥茧，揭示文学的深层内涵；而借助计量方法，通过对作品中某些语言现象的统计分析，可以为文学作品的创作特征及其内涵研究提供科学依据。

本文将以曼德尔施塔姆及赫列勃尼科夫这两位诗人为例，借助俄罗斯国家语料库的诗歌库，通过对所提取的词汇语义类别进行统计分析，研究诗人的创作特征，以此为例阐述词汇

语义分析在文学计量研究中的作用。

2 语料库的语义标注及语义场选择

2.1 俄罗斯国家语料库的语义标注

2003 年开源的俄罗斯国家语料库（以下简称 НКРЯ），经过 20 年的不断更新迭代，已发展为一个含 20 亿词的大型数据库。语料库的内部构成，除了基础库，还建有针对不同研究目标的报刊库、句法库、口语库、平行库、方言库等，并有多种检索功能。

本论文数据采集所使用的诗歌库，自 2006 年开始建设，至 2008 年中期，建成了包含 48 位诗人，约 200 万词次的数据库（Гришина и др. 2009：73—74）。时至今日，该库已有 18 世纪至今 973 位诗人的 94932 个文本，总计 1300 余万词次[2]。除了与基础库共享的词法及语义标注，诗歌库的研究者还开发出了用于格律研究的诗格标注系统。不过，本研究聚焦在诗歌的内容方面，不涉及格律，因此，主要运用的是其语义场检索功能。НКРЯ 的语义标注体系自上世纪 90 年代开始研制，在词汇语义研究的基础上建立了语料库的语义特征分类系统，迄今为止，该系统已具备所有实词词类的基本语义类别。如动词的语义类别有：движение, помещение объекта, физическое воздействие, бытийная сфера, местонахождение, посессивная сфера, ментальная сфера, эмоция, речь, поведение человека 等；形容词的语义类别有：размер, форма, цвет, вкус, запах, температура, место, время, свойство человека 等。名词分为抽象名词、具体名词、专有名词三大类，抽象名词的语义标注有：мероприятие, болезнь, спорт, игра, единица измерения 等，具体名词为：лица, животные, растения, вещества и материалы, здания и сооружения, инструменты, транспортные средства 等。其中与本文相关的名词词汇语义标注构成如下图：

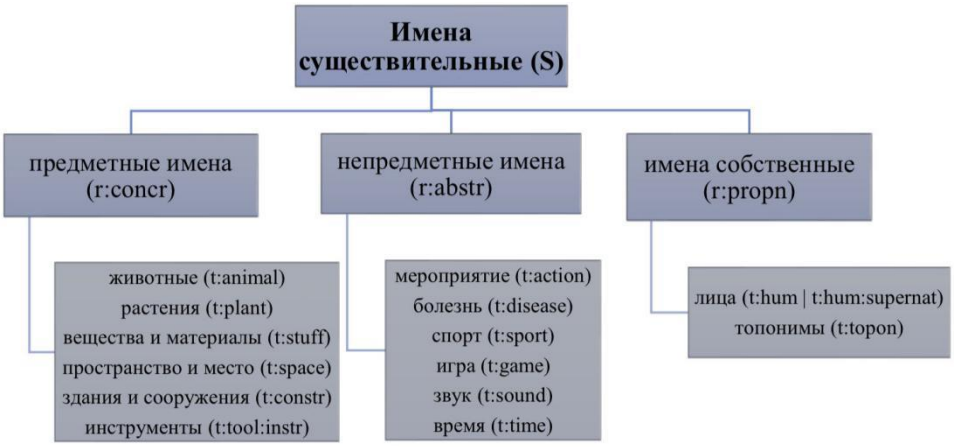

图 1 俄罗斯国家语料库中名词语义层级标注示意图

语料库的语义标注为文本的内容分析提供了极为丰富的数据资源。就名词而言，通过检索作家文本中不同语义类别的词位及词频分布，并对所收集的数据进行统计分析，可以发现或验证其创作特征。

2.2 本研究的语义场选择

语料库丰富的资源常常让研究者无从下手。动词、名词、形容词、副词，以及各词类的语义场，到底选取哪些数据做统计分析？这取决于作者拟研究的问题。对于本论文探讨的诗歌创作而言，可以根据诗人的不同创作特点选取相应的语义场。

作为俄罗斯白银时代阿克梅派诗人的代表，曼德尔施塔姆抱有“对世界文化的眷念”这一诗学理念（阿格诺索夫 2001：234），这种理念无疑体现于其诗作中。鉴于《石头集》“‘浓缩’了诗人艺术世界几乎所有的特点”（阿格诺索夫 2001：236），我们将研究范围锁定为

这部诗集。赫列勃尼科夫则是俄罗斯最重要的未来派诗人之一，是立体未来主义的领袖人物，其早期作品《诗集》“具有原始主义艺术特征”“古斯拉夫文化、古埃及和古希腊罗马文化的汇集，构筑起一个同根同源的世界文化”（王永 2018：678）。那么，这些特点是否贯穿于诗人的创作始终这一问题，尚无定论，须对更多的诗作进行研究。

可以看到，两位诗人的作品中都凝结着作者本人对世界文化的考察。特尼亚诺夫曾指出专有名词在诗歌语言中的重要作用，认为“（专有）名词的词汇色彩作用非常强”“带有专有名词的诗行……具有强烈的诗歌情感”（Тынянов 2018：121，123）。我们在研究中也发现，诗人作品中包含的专有名词，尤其是人名、地名，不仅赋予了作品特殊的色彩，还是多种文化的代表，可以从一个侧面反映诗人的创作特征，乃至体现出诗人的创作理念。因此，本文以人名、地名这两个语义场为数据统计范围，分别对语料库中曼德尔施塔姆的《石头集》以及赫列勃尼科夫的整体诗歌创作进行统计，以此对诗歌的计量研究开展实验。

НКРЯ 诗歌库中收入的曼德尔施塔姆诗作有 679 个文本，计 59587 个词。按人名、地名条件检索，获得 568 个文本，2529 个词。之后，根据《石头集》创作的时间进行筛选，获得所需专有名词词表。以相同方式检索，可提取出赫列勃尼科夫诗作的 290 个文本，53928 个词次；经语义检索，包含人名、地名的诗作有 242 个文本，2256 个词。

不过，俄罗斯国家语料库的语义自动标注未经过清洗处理，错误率比较高，须人工校对。校对后，曼德尔施塔姆《石头集》的人名、地名共计 82 个词位，109 词次；赫列勃尼科夫诗作中的人名、地名共计 193 个词位，378 词次。以下我们分别对两位诗人诗作中含这两类语义的词汇做统计分析，以期揭示诗人的某些创作特征。

3 曼德尔施塔姆《石头集》的人名地名统计分析

3.1 《石头集》人名数据统计分析

曼德尔施塔姆《石头集》中属于 t:hum（人名语义场）的有 50 个词位，57 词次，词频分布如图 2 所示。

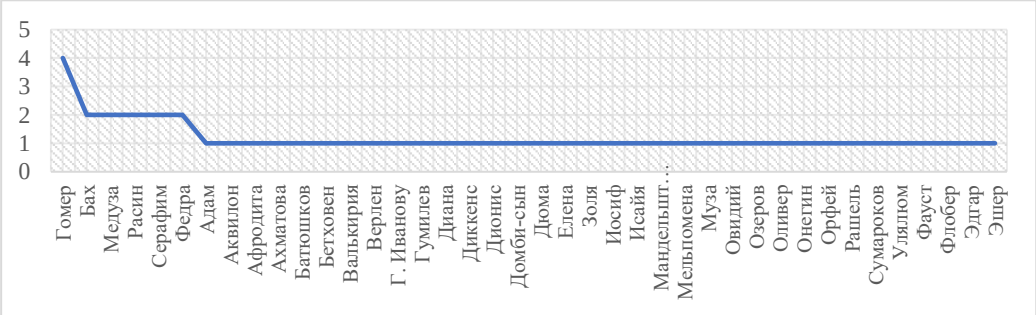

图 2 曼德尔施塔姆《石头集》人名词频分布图

如图 2 所示，人名的词位数与词频数之差很小，体现出较高的词汇丰富度。最高频的词位是荷马。50 个人名中，既有真实人物，又有虚构人物。我们按人物所处的时间分类列表（表 1），可以清晰呈现人名的时间跨度。鉴于历史分期与文化史分期不完全吻合，列表根据本文研究目标，以欧洲文化史为分期。真实人物直接按其所处阶段分类，虚构人物则按作品产生时间分类；其中《圣经》人物涉及《旧约》和《新约》，最早成书于古代，但中世纪盛行教会文化，因此，列表中的《圣经》人名贯穿两个时期。

表 1 曼德尔施塔姆《石头集》人名列表

文化史时期	真实人物	虚构人物	
古典时期	Сократ, Гомер, Цезарь, Цицерон, Август, Овидий, Юстиниан	Аквилон, Афродита, Диана, Дионис, Елена, Мельпомена, Муза, Орфей; Валькирия	Адам, Иосиф,

中世纪时期			Исайя, Серафим
近代时期	Батюшков, Бах, Бетховен, Бонапарт, Диккенс, Дюма, Золя, Лютер, Меттерних, Озеров, Петр, Расин, Сумароков, Флобер, Эдгар	Домби-сын, Евгений, Оливер, Онегин, Улялюм, Фауст, Федра, Эшер	
现代时期	Ахматова, Бенедикта XV, Верлен, Г. Гумилев, Иванов, Мандельштам, Рашель		

从表 1 可以看出，《石头集》的人名，除了未见文艺复兴时期的人物，其余人物涵盖了欧洲文化史自古到今的跨度。如古典时期的苏格拉底、荷马、奥古斯都，中世纪时期的《圣经》人物，近代时期的马丁·路德、拿破仑、彼得一世、狄更斯、福楼拜、左拉，现代时期的魏尔伦、阿赫玛托娃、本笃十五世等。

根据人名所属性质，还可以归纳出以下特征：

从人物性质上看，有君主帝王（查士丁尼大帝、凯撒、彼得大帝、拿破仑）、哲学家（苏格拉底）、政治活动家（西塞罗）、宗教改革家（马丁·路德）、文学家（荷马、奥维德、拉辛、狄更斯、爱伦·坡、阿赫玛托娃等）、音乐家（巴赫、贝多芬等）、神话人物（狄奥尼索斯、海伦、狄安娜等）、《圣经》人物（亚当、约瑟夫、以赛亚等）、文学艺术作品人物（浮士德、费德拉、尤娜路姆等）。

从虚构人物所属的文学艺术流派上看，诗集的人物几乎包含了欧洲文学与艺术史的主要流派：古希腊罗马神话、《圣经》故事、巴洛克风格（作曲家巴赫）、古典主义（作曲家贝多芬；文学家拉辛、苏马罗科夫、奥泽罗夫等）、浪漫主义（作家爱伦·坡）、现实主义（作家狄更斯、福楼拜）、自然主义（左拉）、现代主义（魏尔伦、阿赫玛托娃）。

从人物所属国别上看，主要集中于欧洲国家，分别为古希腊（荷马）、古罗马（奥维德）、法国（路易、拿破仑、拉辛、福楼拜、左拉）、俄罗斯（彼得大帝、苏马罗科夫、奥泽罗夫、巴丘什科夫、阿赫玛托娃、古米廖夫等）、德国（马丁·路德、巴赫、贝多芬）、英国（狄更斯）、美国（爱伦·坡）等。

总体而言，《石头集》的人名涉及社会政治、哲学、文化、文学艺术等领域，涵盖欧洲文化史的主要发展阶段及文学艺术的主要流派，具有跨时代、跨地域及多样性特征。

此外，在上述人名中，有两种类型的人名比例非常高，这一特征尤为突出。其一是同古希腊罗马文化相关的人名，有 16 个，占人名总词位数的 32.00%；其二是文学家艺术家及其作品中的人名，有 40 个，占总词位数的 80.00%。

3.2 《石头集》地名数据统计分析

《石头集》中属于 t:topon（地名语义场）的名词共有 32 个词位，52 词次。词频分布如图 3。

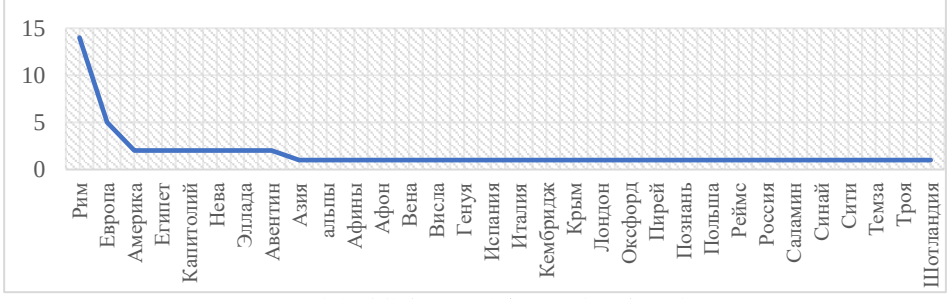

图 3 曼德尔施塔姆《石头集》地名词频分布图

如图 3 的数据显示，最高频的地名是罗马，有 14 频次；其次是欧洲，5 频次。这些地名区域分布特征非常明显，按其所属大洲、国家、行政区划、地形地貌等类别划分，可以列出表 2。

表 2 曼德尔施塔姆《石头集》地名列表

大洲	国家	行政区划	地形地貌
Азия			
Америка			
Европа	Испания, Италия, Польша, Россия, Эллада	Афины, Вена, Висла, Генуя, Кембридж, Крым, Лондон, Оксфорд, Пирей, Познань, Реймс, Рим, Троя, Шотландия	Авентин, Альпы, Афон, Капитолий, Нева, Саламин, Сити, Темза
（非洲）[3]	Египет		Синай

将表 2 分类与其数据结合分析，可以发现，《石头集》中的地名涉及亚洲、美洲、欧洲和非洲的国家及城市、河流、山川等名称。其中欧洲地名有 28 个，占地名总词位数的 87.5%。由此构成了一个以欧洲为中心，辐射至其他地域的空间网络。

其次，在这些地名中，同古希腊罗马文化相关的有：国家名称爱拉多（即古希腊）；城市名称罗马、雅典、比雷埃夫斯；山丘名称阿芬丁山、卡皮托利山、阿索斯圣山；岛屿名称萨拉米斯岛；《荷马史诗》中的特洛伊城等。这些词共 9 个，占总词位数的 28.13%。

从词频分布的情况看，出现频率最高的国家是意大利，有 4 个词，19 频次，占地名总词频数的 36.54%。其中罗马一词出现 14 次，占地名总词频数的 26.92%，是诗集中词频最高的专有名词。

3.3 结论

基于以上数据分析，可以得出结论：

1）《石头集》体现出阿克梅派的特征——“对世界文化的眷念”。诗人笔下的“世界文化”网络，是一个上至古希腊罗马，下至诗人所处时代，以欧洲为中心，辐射到美洲、亚洲和非洲的时空域；

2）在诗人的“世界文化”网络中，古希腊罗马文化占有独特地位，其中罗马构成了“世界文化”的核心；

3）在“世界文化”网络中，文学艺术构成其最为重要的载体。

4 赫列勃尼科夫诗作的人名地名统计分析

4.1 赫列勃尼科夫诗作的人名数据统计分析

国家语料库收录的赫列勃尼科夫诗作中，属于 t:hum（人名语义场）的有 106 个词位，188 词次。词频分布如图 4。

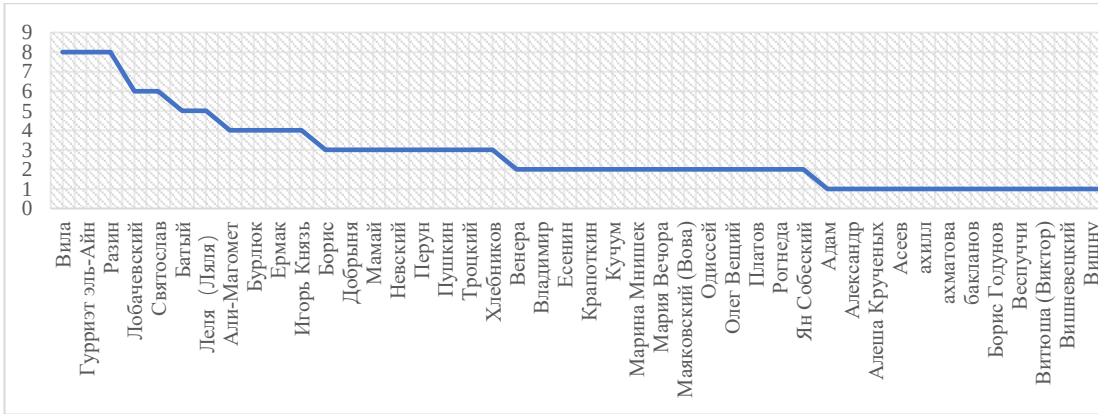

图 4 赫列勃尼科夫诗作人名词频分布图

如图 4 所示，人名中最高频的三个词分别是斯拉夫民族民间传说中的人物维拉、俄国农民起义领袖拉辛，以及波斯女英雄古力艾特-艾力-爱因；其次是数学家罗巴切夫斯基和俄国大公斯维托斯拉夫；此后是与俄国历史相关的拔都汗、古斯拉夫神话人物列利亚；再后是俄国历史人物伊戈尔王和叶尔马克、埃及帕夏穆罕默德、画家布尔柳克。这几个高频词分别是以下五类人物的代表：历史（尤其是俄罗斯历史）、东方文化、古斯拉夫文化、科学、文学艺术。这些类别的人名有 101 个词位，181 词次，占了所有人名词位的 95.28%，词频的 96.28%。我们将这些人名按这五类列表如下（表 3）

表 3 赫列勃尼科夫诗作人名列表

类别	人名
历史	Александр, Бакланов, Батый, Борис Годунов, Вишневецкий, Владимир, Волконский, Врангель, Глеб, Грозный, Дибич, Добрыня, Ермак, Ермолов, Игорь Князь, Кутузов, Леля, Мамай, Марина Мнишек, Монтезум, Невский, Олег Вещий, Ольга, Ослябя, Остраница, Платов, Пугачев, Разин, Рогнеда, Романов, Святослав, Станислав, Суворов, Урсула, Ян Собеский
东方文化	Али-Магомет, Вишну, Гурриэт эль-Айн, Заратустра, Пржевальский, Суэ
古斯拉夫文化	Вила, Гуль-мулла, Леший, Морозенко, Перун, Садко
科学	Веспуччи, Галилей, Лобачевский, Нансен
文学艺术	Бурлюк, Пушкин, Есенин, Хлебников, Кольцов, Ахматова, Алеша Крученых, Каменский, Лермонтов, Маяковский, Мережковский, Толстой, Тургенев, Тютчев, Асеев, Моцарт, Чайковский, Татлин, Шаляпин

从列表可以看出，赫列勃尼科夫诗作中的历史人物，最多的是俄国历代君王。其中又以古罗斯时期的王公最多，如伊戈尔、涅夫斯基、斯维托斯拉夫、弗拉基米尔、奥列格、奥尔迦、格列布；其他君王有鲍里斯•戈都诺夫、沙皇伊凡雷帝；俄罗斯帝国时期的罗曼诺夫王朝等。与俄国历史相关的还有伪德米特里一世家室成员：玛丽娜•姆尼舍克皇后和她的哥哥斯坦尼斯拉夫、姐姐乌尔苏拉、姐夫维什涅维茨基；与蒙古人统治时期有关的拔都汗、马迈汗、库楚姆汗、库里科沃战役中的英雄奥斯利亚比亚神父；武士多布雷尼亚；各种战争中的统帅，如军事家季比奇、巴克拉诺夫、沃尔康斯基、库图佐夫、符朗格尔等，以及哥萨克首领普拉托夫、奥斯特拉尼察；俄国农民起义领袖拉辛和普加乔夫，远征西伯利亚的叶尔马克。涉及其他国家历史的人物有埃及阿里王朝的建立者穆罕默德•阿里帕夏，马其顿国王亚历山大大帝，波兰国王扬三世•索别斯基，墨西哥印第安族统治者蒙特祖玛。

令人引起东方文化联想的人名有波斯历史人物查拉图斯特拉，波斯女英雄古力艾特-艾力-爱因印度婆罗门教的保护神毗湿奴，印第安人的太阳神苏艾，俄罗斯旅行家及科学家普尔热瓦尔斯基。后者之所以与东方文化有关，主要由于他曾多次考察中亚地区，到过蒙古和中国。

同古斯拉夫文化相关的有古斯拉夫神话人物列利亚、雷神皮隆、林妖，民间传说中的英雄萨特阔、莫罗赞柯等。

科学领域的人名，有非欧几何发现者、俄国数学家罗巴切夫斯基，意大利科学家伽利略，物理海洋学的奠基人、挪威探险家和科学家南森，意大利航海家和探险家韦斯普奇。

文学家和艺术家的人名中，既有 18—19 世纪经典文学家和音乐家，又有赫列勃尼科夫本人及其同时代人。前者如：柯里佐夫、普希金、莱蒙托夫、屠格涅夫、丘特切夫、托尔斯泰等文学家，以及莫扎特、柴可夫斯基等音乐家；后者如梅列日科夫斯基、阿赫玛托娃、阿谢耶夫、克鲁乔内赫、卡缅斯基、马雅科夫斯基、叶赛宁等文学家，以及布尔柳克、塔特林、夏里亚宾等艺术家。

由此可见，赫列勃尼科夫诗作中的人名，涉及“‘各个时代的人物’和‘各种文化的参与者’”（俄罗斯科学院 2006：166），体现出诗人对古老历史、东方文化以及古斯拉夫文

化的关注与思考。

4.2 赫列勃尼科夫诗作的地名数据统计分析

国家语料库收录的赫列勃尼科夫诗作中，属于 t:topon（地名语义场）的名词有 87 个词位，190 词次。词频分布如图 5。

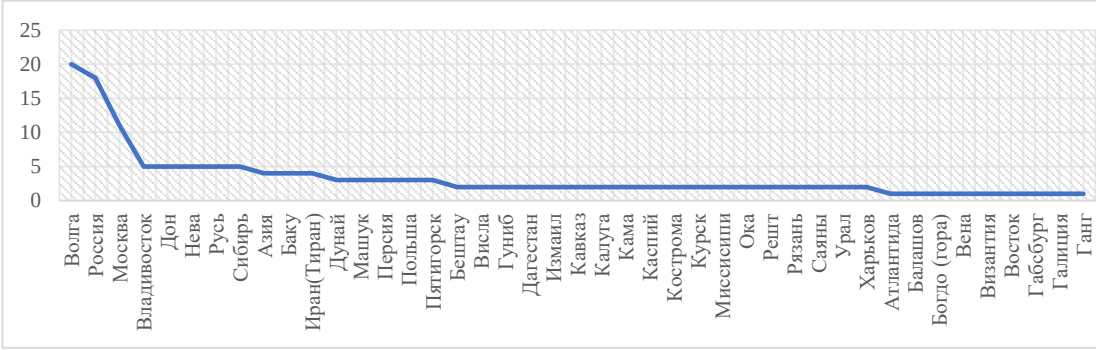

图 5 赫列勃尼科夫诗作地名词频分布图

如图 5 所示，诗作的地名中，最高频的是伏尔加河，有 20 词次；其次是俄罗斯，18 词次（如果加上俄罗斯的古称“罗斯”，则为 23 词次，跃居首位）；莫斯科 11 词次。因此可以说，赫列勃尼科夫是非常“俄罗斯”的诗人，是“伏尔加河”之子。按所属大洲、国家、行政区划、地形地貌等类别分类，可以列出表 4。

表 4 赫列勃尼科夫诗作地名列表

区域	国家	行政区划	地形地貌
（欧亚）	Россия, Русь	Баку, Балашов, Владивосток, Гуниб, Дубравный, Златоуст, Измаил, Казань, Калуга, Каргебиль, Кисловодск, Китеж, Кострома, Красноводск, Москва, Наргинь, Петровск, Поле Девичье, Полтава, Пятигорск, Рязань, Самара, Севастополь, Сибирь, Сочи, Судак, Тула, Тучков, Углич, Хаджи-Тархан, Харьков, Царицын	Волга, Днепр, Дон, Иртыш, Кама, Каспий, Нева, Неман, Ока, Хопер; Бештау, Богдо, Дубравный, Жучка, Кавказ, Казбек, Курган Золотой, Машук, Саян, Урал
（欧洲）	Габсбург, Гогенцоллерны, Греция	Вена, Висла, Галиция, Мюнхен, Нарва, Перемышль, Плевна, Польша, Шипка	
Азия	Индия, Иран, Персия, Турция, Япония	Византия, Пекин, Решт, Царьград	Ганг
（美洲）		Чикаго	Миссисипи
（非洲）	Египет		Нил

地名列表及统计数据显示，从地域分布上看，位于俄罗斯（含俄国、苏联）的地名最多，有 58 个词位，155 词次；其次是涉及亚洲的地名，有 10 个，19 词次；其他欧洲国家的地名，有 12 个词位，15 个词次；此外，还有非洲、美洲的地名，如埃及、尼罗河、芝加哥、密西西比河等。就俄罗斯本国的地名而言，有几个特征比较突出：

1）河流名称分布：不仅有高频词伏尔加河，还有卡马河及奥卡河、顿河及多瑙河。前者构成了伏尔加河的支流，后者则令人联想到位于期间的东欧大草原，那里曾经是东伊朗语族的游牧民族西徐亚人生活的地域。

2）山川名称分布：除了高加索和乌拉尔这两个词本身，还有皮亚季戈尔斯克市周边的马舒克山、别什塔乌山、卡兹别克山、茂林山、金色山丘；还有位于西伯利亚的萨彦岭。

3）城市（含岛屿）名称分布：城市名最高频的是莫斯科；其次是符拉迪沃斯托克和西

伯利亚。还有伏尔加河沿岸的科斯特罗马、萨马拉、乌格利奇、察里津（现伏尔加格勒）、哈吉-塔尔汗（现阿斯特拉罕）；里海之滨的巴库、克拉斯诺沃茨克[4]，岛屿纳尔金；黑海之滨的伊兹梅尔、索契；克里米亚的塞瓦斯托波尔、苏达克；高加索地区的皮亚季戈尔斯克、基斯洛沃茨克；村庄古尼勃、卡尔格比利；乌拉尔地区的兹拉托乌斯特等。

由此可见，赫列勃尼科夫笔下的地名，勾画出一个以伏尔加河为中心线，北至莫斯科，南达阿斯特拉罕，向西向东辐射，西至西欧、东到远东，南部延展至黑海和里海的辽阔区域。在此区域中，高加索、乌拉尔山脉以及伏尔加河流域，成为地名分布的核心。这种特征不仅与赫列勃尼科夫的生平及创作经历有关，更显示出诗人对俄国发展史的关注，对自己家乡与俄国历史关系的思考。

赫列勃尼科夫出生在阿斯特拉罕省切尔诺亚尔斯克县的杰尔别特村，全长 3692 公里的伏尔加河流到这里注入里海，西南为高加索山脉，东北为乌拉尔河及乌拉尔山脉，处于欧亚交界的平原上。这里曾是金帐汗国的一部分，在 1466—1556 年间是阿斯特拉罕汗国的都城。特殊的地理位置使得诗人格外关注这片土地上的历史沿革，考察家乡的历史，考察俄罗斯与东方文化之间的联系。

4.3 结论

基于以上数据分析，可以得出结论：

1）赫列勃尼科夫诗作中的人名地名，构成了多元世界文化叙事的一个又一个节点；

2）俄罗斯、伏尔加河、莫斯科构成了诗人笔下艺术世界的中心，而与古斯拉夫及东方相关的历史和文化人名地名，构成了由中心向外辐射的叙事载体；

3）在这个多元世界文化中，诗人对历史、东方文化和古斯拉夫文化情有独钟。

5 结语

综上所述，可以认为，词汇语义统计分析对文学研究具有其独特的价值。首先，人名、地名的数据统计分析可以论证前人研究中得出的某些结论，比如证实了曼德尔施塔姆创作显示出“对世界文化的眷念”，赫列勃尼科夫的创作具有“与科学、文化、历史等各个领域无限广泛的联系”（阿格诺索夫 2001：376）等；其次，数据统计分析可以清晰地揭示出诗人创作最突出的特征，比如曼德尔施塔姆的诗歌体现出世界文化以欧洲为中心，而赫列勃尼科夫的诗作则明显倾向于东方；最后，这种研究使得文学研究的结论以科学依据为基础，具有更高的可信度。

运用语料库数据进行统计分析的文学研究看似简单，在实际操作中，会碰到很多具体问题。文学的计量研究有两个必经的重要环节：首先，需要从问题着手，先根据作家的创作特点初步确定研究问题，再考虑解决该研究问题所须提取的数据；其次，由于语料库是计算机对大数据做自动处理的产物，鉴于语义的复杂度，语料库的语义标注准确率相对较低，后期需要人工校对。因此，文学计量研究的难度并不比定性研究小，但这种将定量与定性相结合的研究范式，可以为文学研究开拓新的视角，并为其带来一些新的发现。

附注

1 相关论述参见：王永、李昊天、刘海涛，俄罗斯计量语言学发展述评[J]. 外国语，2017(6).

2 数据统计源自俄罗斯国家语料库：[EB/OL]. https://ruscorpora.ru/, 2023-04-08.

3 地名列表中，括号内文字仅表示类别，并非诗作中出现的词位。

4 巴库和克拉斯诺沃茨克现分别为阿塞拜疆及土库曼斯坦城市，本文按赫列勃尼科夫所处时代的归属划分。

参考文献

[1]Гришина Е. А., Корчагин К. М., Плунгян В. А., Сичинава Д. В. Поэтический корпус в рамках НКРЯ: общая структура и перспективы использования[A]. Плунгян В. А. Национальный корпус русского языка: 2006-2008. Новые результаты и перспективы[C]. СПб.: Нестор-История, 2009.

[2]Пенковский А. Б. Очерки по русской семантике [M]. Москва: Языки славянской культуры, 2004.

[3]Тынянов Ю. Н. Проблема стихотворного языка [M]. Москва: КомКнига, 2018.

[4]弗•阿格诺索夫主编. 白银时代俄国文学[M]. 石国雄、王加兴译. 南京：译林出版社，2001.

[5]俄罗斯科学院高尔基世界文学研究所集体编写. 俄罗斯白银时代文学史（IV） [M]. 谷羽、王亚民等译. 兰州：敦煌文艺出版社，2006.

[6]王　永. 外国文学的计量研究——研究背景、发展现状及研究路径[J]. Interdisciplinary Studies of Literature，2021(4).

[7]王　永，黄锦南. 赫列勃尼科夫《诗集》的原始主义艺术特征[J]. Interdisciplinary Studies of Literature，2018(4).

Application of Lexical Semantic Analysis in Literary Quantitative Research

Wang Yong

(Zhejiang University, Hangzhou 310058, China)

Abstract: Statistical analysis has been used in literary research in Russia for over a century, with a particular focus on areas such as poetic metrics, authorial style, and author identification. However, the study of Russian literature in China is still in its early stages. This paper seeks to contribute to this field by applying data statistical analysis to the study of poetry. The Russian National Corpus serves as our data source, with lexical semantic features as the retrieval entry. By analyzing the data of personal and geographical names in the poetry texts of Mandelstam and Khlebnikov, statistical analysis is conducted on the lexical semantic features of poetry texts. The results demonstrate that literary text analysis based on lexical semantic statistical analysis can help reveal the creative characteristics of works and the author's creative themes.

Keywords: lexical semantic categories; poetry; quantitative research

基金项目：本文系国家社科基金重大项目“中国外国文学研究索引（CFLSI）的研制与运用”（18ZDA284）的阶段性成果。

作者简介：王永（1965—），浙江杭州人，浙江大学外国语学院教授，博士生导师，研究方向：俄罗斯文学、词汇语义学。

收稿日期：2023-04-25　　[**责任编辑：信　娜**]