《18—19 世纪上半叶俄语词典合集》的数字化开发及其对词汇历时研究的作用

石　琦

（黑龙江大学俄罗斯语言文学与文化研究中心，哈尔滨 150080）

提　要：本文聚焦俄罗斯喀山联邦大学开发的《18—19 世纪上半叶俄语词典合集》，介绍其对所收录的词典词条文本的加工及检索功能的设置，探讨该合集在俄语历史词汇学研究中的重要作用，并结合数字化时代融媒词典的发展趋势提出了有关该合集进一步完善的建议。

关键词：《18—19 世纪上半叶俄语词典合集》；数字化；词汇的历时研究

中图分类号：H356　　**文献标识码：A**

1 引言

俄罗斯词典学家莫尔科夫金（В. В. Морковкин)在 40 多年前提出了“词典体系”（система словарей）这一概念，该体系“应包含各个类型的词典，对同一对象（同一语言单位）从不同角度加以描写”（Морковкин 1986：106），体系内的多部词典各司其职、相互补充，成为拥有统一概念、相同理论方法和研究取向的辞书作品合集，扩展用户对于语言及语言词汇规模的认知。这一概念在诞生之初被认为是理想化的抽象概念，需要国家层面上的长期远景的规划。进入 21 世纪以来，全球科技创新进入前所未有的密集活跃期，以网络化、信息化、智能化为支撑的数字化转型已成大势所趋，“数字化技术不仅改变了大众传媒获得信息的时间、空间及其成本，更主要的是为电信业、出版业和广播电视业出现产业融合提供了重要的技术支持”（章宜华 2019：81）。当代词典也紧随这一时代浪潮“由平面媒体向多媒体、多模态和融媒体方向发展”（章宜华 2021a：21）。这为词典的数字化开发带来了机遇，使得“词典体系”这一概念成为可能。喀山联邦大学列夫·托尔斯泰语言学和跨文化交流研究所依托这一理念，提出整合 18—19 世纪初各类词典资源形成综合性的词典合集（Комплексный справочный фонд словарей русского языка XVIII — первой половины XIX века）（以下简称《合集》）的项目，该项目模糊词典类型的概念，将各类词典资源集群化，《合集》内收录了包括《俄罗斯学院词典》（Словарь Академии Российской，以下简称《学院词典》）、《按字母顺序排列的俄罗斯学院词典》（Словарь Академии Российской, по азбучному порядку расположенный）和《教会斯拉夫语与俄语词典》（Словарь церковнославянского и русского языка）在内的多部划时代经典辞书作品，凝结了 18—19 世纪初众多语言学家的智慧结晶，反映了这一时期的社会思想和俄语使用者的集体语言意识。《合集》公布于喀山联邦大学“喀山语言学基金会”（Казанский лингвографический фонд）门户网站（http://www.klf.ksu.ru）上，作为一种互联网信息资源，供所有网络用户免费访问。本文从该合集的文本加工、检索方式以及其对词汇历时研究的作用三个方面对其进行介绍，

以期使我国词典学及词汇学领域的研究者了解俄罗斯当代数字词典词典编纂的新动态。

2 《合集》的文本加工

“数字化媒体融合是辞书产业的必然趋势。词典与融媒体结合涉及文本组织形式、文本传播媒介和词典文本的使用三个部分，但其核心仍是词典文本。因此，传统纸质词典的文本内容仍有很大现实价值”（章宜华 2022：13）。《合集》十分注重词典文本的整合加工，没有完全照搬词典古籍的原始文本，而是对其进行了必要的加工与改进，力求提升词条信息的质量，为用户理解词汇语义、语法及语用环境提供更为全面的参考。

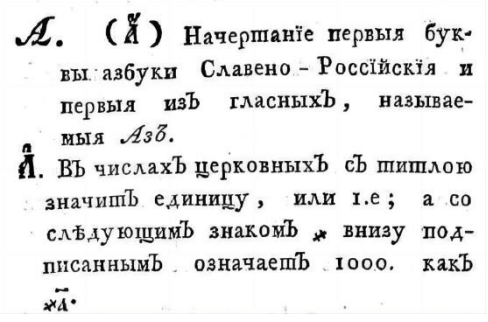
2.1 词条信息勘误

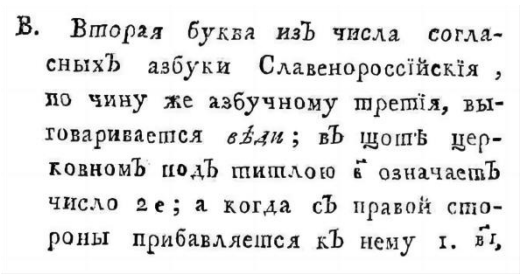
《合集》收录的多部词典编纂于民族标准语形成的过程之中，在规范俄语标准语方面发挥了积极作用。然而由于语言规范正处于发展的过程之中，因此存在很多有争议的情况，加之词典编纂工作千头万绪、纷繁复杂，难免百密一疏，为了避免词条信息数字化转化过程中出现“继承性”错误，尽可能保障词条内容的准确性，《合集》秉持着仔细查证、认真核验的原则修正了原版词典中错误的内容，尤其是词目的拼写错误。例如，在对《俄罗斯学院词典》文本进行加工的过程中，将 вбаживаю 改正为 вваживаю，将 муроварица 改正为 мурoварница 等，此类拼写错误通常较为明显，可以通过词族内其它词条中的词汇材料得以证实。同时《合集》也参考了其它词典的相关词条信息，如《教会斯拉夫语与俄语词典》、《18 世纪俄语词典》（«Словарь русского языка XVIII века»）等权威词典；相关记录可在《合集》参考资料区（справочная зона）查询，各类更正信息可在《合集》词条注释中查阅。《学院词典》是一部规范性词典，是之后多部词典所参考的蓝本。《合集》重新审核其词条信息，提高了词汇形态的可靠度，也提升了词条信息的整体质量；将此类加工后的多部词典的更为优质精确的词条信息协调整合，可构成权威的历史文本数据库，供读者查考词汇语源、追溯词形演变；此类大规模的权威数据库又可为今后的词典编纂提供知识数据，有着极为珍贵的学术价值。

2.2 词条信息数据的精细化切分

《合集》力求对词汇的语言学信息进行全面的描写，仍以《学院词典》的文本加工为例，《合集》提取了《学院词典》中词汇的词素、熟语、谚语等信息，且对各信息项进行了切分与分级处理，使得数据结构更为严密、内容层级更加分明。各数据项既是相对独立的，可以根据用户需求提供碎片化的精准服务；又是互相联结的，可融合为整个词条，系统化呈现各类信息。于是各类数据切分的精细化程度直接决定了《合集》的智能程度，《学院词典》文本的提取过程中尤为注重这一点，在处理文本的过程中根据各数据项的特征对词条信息进行了详细具体的标注。比如，《合集》逐条对原版词典中的同形异义词进行了细致的区分并予以显著的标识（置于括号中），帮助读者比较分析检索词，进行词条的筛选。根据同形异义词的不同类型提供词类、重音、词义、上位词等一系列标识，例如：алътъ（嗓音）和 алътъ（小提琴），возъ（名词）和 возъ（前置词），навозить（о 为重音）和 навозить（и 为重音），避免由于同形异义词存在而产生的语义混淆和词汇误用，提高查词准确度，提升检索效率。

再者，《合集》改善了原版词典内同类信息描写方式不统一的情况，例如，原版词典对于“А”和“В”这两个词条的描写内容基本一致，却在编排体例上参差不一（见图 1）。


А. (А) Начертаніе первыя буквы азбуки Славено-Россійскія и первыя изъ гласныхъ, называемыя Азъ.
А. Въ числахъ церковныхъ съ титлою значитъ единицу, или 1.е; а слѣдующимъ знакомъ ҂ внизу подписаннымъ означаетъ 1000. какъ А҂.


В. Вторая буква изъ числа согласныхъ азбуки Славенороссійскія, по чину же азбучному третія, выговаривается вѣди; въ щотѣ церковномъ подъ титлою в означаетъ число 2е; а когда съ правой стороны прибавляется къ нему 1. в҃і,

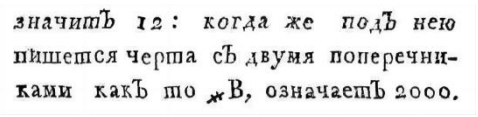

значитъ 12: когда же подъ нею
пишется черта съ двумя поперечни-
ками какъ то ж̃В, означаетъ 2000.

图 1 《学院词典》中的“A”和“B”词条

很明显，“A”对同类信息的描写收录在两个词条中，“B”中的同类信息则是在一个词条中给出。而《合集》在提取词条信息时对于此类相似的条目进行了比较分析，按照各类信息所属范畴构建分类模型，对相似的条目信息按照统一的释义模式进行处理，形成层次清楚、组织有序的词条结构，力求达到信息间、词条间的相互融通。

值得一提的是，《合集》内的某些词条信息已深入到词素层面，例如 без 一词的词条内提到了作为前缀的 без-的语义内容，词素层面的语义描写将以 без-为前缀的词汇联结起来，形成一个以词素为枢纽的词汇网络，促进了相关语词词义的理解。

2.3 词条信息数据的完善与联结

传统纸质词典由于受到规模限制，在词典结构信息项的编排上都要考虑到结构与篇幅问题，而《合集》作为一个综合性的电子信息系统则“无容量限制，在编排和呈现形式上更强调易查、易得、易懂，”（章宜华 2021b：105）。所以我们可以将整理词条信息的过程看作是对于原版词典的再一次“创编”，通过数字化技术手段对词条进行系统化的整理与比较，提取词典信息并按照既定的分类组织储存为相对独立又与词目词相关联的数据化的词典信息项。

首先，《合集》根据词条内容将纸版词典囿于规模限制未收录的词汇补充为条目词，一是在纸版词典中以“参见”形式提及的词汇，例如《学院词典》中 аангичъ 词条内提到：“……在西伯利亚，它被称为 Савка 或 Саутка”，其中的 савка 一词在词典中是一个独立的条目，而 саутка 却没有被收录为条目词，只是在 аангичъ 和 савка 的词条中被提及，《合集》将 саутка 作为条目词补充收录。二是与纸版词典中的条目词有形态（词族）关联的词汇，例如，《合集》考虑到已收录的词汇短尾形容词 близехонекъ 及副词 близехонько，将《学院词典》中未提及的长尾形容词 близехонъкий 也增添收录入词表内，加筑了语词间的关系网络。《合集》没有了纸质辞书规模的限制，也不再受宏观结构中词汇排序的制约，因此在编排上倾向于方便用户一键通查，利用结构化的数据将有语义关联的词汇联结起来，强化了中观结构的桥梁纽带作用，帮助用户构建清晰的词汇信息关联网络，提升了用户查询的效率和使用词典的体验。

其次，《合集》将纸版词典的词条信息重新编码整理，将词条中作为例证的短语、谚语、俗语作为该语段中所包含的其他条目词的同类数据储存，构成词条数据的关系网络。用户在查询词汇时，既可以检索词目词，又可以通过短语等复杂字段进行查询，例如，谚语“На смирнаго Богъ нанесешь, а резвой самъ набежитъ”在纸质版《学院词典》中只收录于该条谚语的中心词 набегать（набегаю）的词条中，而在《合集》中用户可通过查询谚语中其他的词汇单位或语段得到该条谚语的相关信息，此种各词条之间数据的共享互通极大地拓展了《合集》的信息利用潜力。

总而言之，《合集》将语料重新整理联结的过程就是一次将“半成品”拆分开、添加必要的“部件”并将这些零部件重新组合为“成品”的过程，这一新的“成品”冲破了传统纸质词典二维平面结构的束缚，以立体的网络化结构组织构建了 18—19 世纪上半叶编纂的各类词典的词条信息，足以保证用户的个性化查阅需求，同时又为俄语历时词典学的研究贡献了一份珍贵的语料，推进了历史文化遗产的保护与传承。

3 《合集》的检索功能

“从有词典设计和词典编纂开始，词典编者就一直潜心研究如何使词典查阅更便宜、更直接”（雍和明 2003：105）。计算机检索系统借助数字化优势，为电子词典检索功能的发展提供了十分便利的条件，打破了传统词典检索中单一的线性检索模式，结合了符号学、信息学、心理语言学等学科的相关概念，“不断朝智能化、多元化、个性化、层次化方向发展，从而实现与数字化潮流、大数据环境的高度兼容”。（王怿旦，张雪梅 2016：42）

3.1《合集》的特色检索方式

《合集》内汇集的词条信息全部取自于 18—19 世纪出版的词典，近三百年的时间跨度对词典检索系统的开发提出了新的要求——即以方便广大读者快速精准查阅为基础，全面展示 18—19 世纪初语言学研究的优秀成果，帮助读者以历时性角度系统了解语言的变化、发展与应用。

首先，考虑到当时通行的俄文字母与现在使用的字母有所不同的情况，《合集》向用户推出使用 18 世纪通行字母索引或是利用现代字母替代停用字母两种查检方式。例如 базар（18 世纪词形为 базаръ），векъ（вѣкъ），миро（мvро）等。对于那些词形发生演变的词汇，系统会对输入词形进行预处理，自动识别过滤词汇的现代变体，将其转换至原本的书写形式。例如 акула（аккула），акциз（акцызъ），олифа（алифа），весьма（веема）等。

此外，《合集》发挥数字化辞书的技术优势，推出了多种创新且实用的检索方式：一是通过直接键入词汇的方式检索。纸质版《学院词典》曾因按词族排列词汇的方式在查检功能上遭受质疑，如今《合集》将其词条信息标注并存储在数据库中，配合精确的索引技术帮助用户通过搜索框检索所需信息，用户只需逐字键入即可直接访问词典数据。二是借助符号检索。用户在不明确词目正确拼写方式的情况下，可采取借助计算机通配符代替若干字母的方式在《合集》检索系统中进行模糊搜索，例如通过查询“без*ный”即可检索到类似于 безумный 的一系列词语；该种方式也可以用来检索相同构式的词组，通过查询«*-*»可得到带有连字符的词语汇总。这种搜索方式十分有助于用户对相同词根或相同词缀单词进行系统归类和比较分析：通过查询«!ать»可检索到一系列以-ать 结尾的四个字母的单词，用户利用符号配合字母输入，即可查阅在构词形态上有联系的一系列词汇，再根据具体需求整合、筛选信息。三是通过输入字母或音节的数量检索。即以词汇最为显著的特性——长度为切入点查检词汇，这一功能为研究人员就词长探究词汇的语言属性、或是就词汇音节数量展开系统的研究分析带来极大便利。

3.2 《合集》的多种检索路径

《合集》利用数字化辞书依托电子信息技术生成检索命令的特点，开辟了多个索引项目供用户自主选择，支持用户设置对释义、词类标注、功能语体标注、词源信息等各类词条数据进行单条件检索或组合检索。用户只需要输入词或语段，再配合设置查询条件，就可以检索到所有包含该词的释义或例证等。如果想要通过词类、词源等信息遴选词汇，则需要借助组合检索提高检索效率，用户根据自身需求设置个性化的索引项目，后台接收索引命令后筛选储存在数据库中的数据项，精确调取所有符合条件的词汇。查检词汇的种种路径为俄语学习者与研究者深入探究词汇信息打开了方便之门。

4 《合集》对俄语词汇历时研究的作用

词汇的历时研究需要依托真实的历史语料，词典作为收录词汇信息的工具书本身就具备文献性质，其中的语料信息高度精炼且具有权威性。《合集》利用现代信息技术（如语料库、数据库等）将古籍词典的词条文本重新切分、提取、标注、整合，使得词条信息结构超越了平面范式，更便于读者与研究人员深层次地探析现代语词的产生与发展历程，通过词汇层面的相关信息深入体会语言中富含的民族精神内核，最终综合多方信息梳理词汇与义项的发展脉络，探究其历史演变。

4.1 词汇历时演变定性研究

“从词汇演化的研究角度来说，历代自然口语和真实文本最能真实而直观地反映词汇的使用情况”（李斌，刘雪扬 2018：153）。《合集》中收入的词典是 18—19 世纪初编写的、经过足够长时间的历史检验的多部科学性经典著作，其中记录了大量真实的历史语料，反映了那一时期词汇使用的实际状况，这十分符合人们对于历史词典的理解，即“展示一种语言某一历史时期词汇状况的词典”（郑述谱 1997：106）。系统性地统计现存精品古籍词典中的词汇信息，加工为能够反映某一历史时期词汇演变的语料库，可供研究人员探究词汇在某一历史时期的真实面貌，也便于从历时的角度观察词汇与义项的产生与消亡。可以说，《合集》搭建了一个运行便利的查阅平台，为词汇历时研究提供了真实可靠的丰富资料。

18—19 世纪初正是俄罗斯民族语言形成的关键期，《合集》收录的多部词典都具备这一时期鲜明的时代特色，例如《学院词典》的收词对象“斯拉夫俄语”十分准确地概括了俄罗斯民族语言在 18 世纪的特定存在形式；词典内部又将词汇按词族排列，一个个相对完整的词族架构清晰地展现了各条目词之间的派生关系，十分便于读者结合相关历史文化对词汇与义项的发展与演变进行系统性定性研究。如今《合集》内经过整合的数字化文本更加便于研究者深入词汇内部从不同角度挖掘所需信息，穿越时空的屏障以现代视角解读词典文本，结合民族文化背景探究俄罗斯民族标准语的起源并分析其历史演变，将 18 世纪多样性的文明继续传承，又以今人的视角赋予其新的解读。

4.2 词汇历时演变计量研究

“俄罗斯的计量语言学起始于 19 世纪中叶，在诗歌节律、词汇系统、文体测量等领域均有深入研究”（王永，李昊天，刘海涛 2017：95），但由于词典文本信息的巨大体量和技术手段的局限，词典词汇计量的工作难度大大提升，例如《合集》中选用的《学院词典》采用了按词族排列词汇的方式，“围绕一个原始词，将所有同根词构建为一个词族”，词族内的层次结构较为复杂，各位学者统计出的词汇总量都不尽相同，现有的统计有超 4 万词（Белоусова 1998：563）、43257 词（Лутовинова 2009：101），不少于 55000 词（Биржакова 2001：109）等几种不同的数据。相关准确数据的缺失使得对于《学院词典》的研究一直集中于定性的描写上，如今《合集》采用数据库技术实现了大量复杂信息的储存、管理和共享，其精细化程度可允许人们对于多部词典进行大规模的定量统计分析，不仅可以统计出每部词典的词汇总量，还可得出各词类词汇占比，单义词与多义词的分布，各长度词汇数量，词典例证中自撰例与书证的比例，归属于不同功能语体的词汇数量等信息，辅助研究人员结合统计方法定量分析 18—19 世纪初词汇的发展状况，对于词汇的考察不再是针对某个词或某类词，而是运用真实权威的数据进行历时维度的整体定量分析。

《合集》提取信息和整合数据的方式也可以为喀山语言学基金会项目任务中其余按年代划分的词典资源合集（如《16—17 世纪文献语言词典》（Словарь языка памятников XVI–XVII вв.）、《17 世纪末至 18 世纪上半叶俄罗斯谚语格言词典》（Словарь языка русских пословиц и поговорок конца XVII — первой половины XVIII в.）等提供参照，按时序逐步填充完善数据库的基础资源，建立词汇历时演变研究的资源性平台——词汇历时检索数据库，从而观测词汇在历代演变的基本情况，梳理词汇语义发展的基本脉络，验证并补充现有的词汇理论与研究成果。

5 思考与启示

《合集》收录多部俄罗斯 18—19 世纪经典辞书作品，以数字化技术为依托，借助新技术、新手段、新理念赋予了古籍辞书新的生命，形成综合性大型古籍辞书数据库，在满足用户快速精准查阅需求的同时，深度挖掘古籍辞书的学术潜力，推进历史文化遗产的保护与传承，在提高网络词典检索效率的同时实现了语言工具书的全球共享。然而文中列举的《合集》

的基本使用功能也只是当前数字化辞书最基本的组成部分，在融媒辞书理念迅速发展的今天，其在内容呈现、编纂模式、产品架构以及后期维护等方面仍然有待进一步完善。比如：1）在内容呈现方面，可通过图片、音效、动画和视频等多元化的信息呈现形式，改善传统词典的缺陷，逐步实现词条信息的多模态化，在激发查阅兴趣的同时，帮助用户通过多感官参与建立其对于查阅内容的立体认知。2）编纂模式方面，可尝试采用“众源模式”，充分调动用户的积极性并吸纳其参与词典编纂，完善《合集》词条内容，集思广益拓宽编纂思路。然而由于《合集》涉及的语言资料大多属于 18—19 世纪的古籍文本，理解起来具有一定难度，需要编纂团队做好审核员和咨询师的工作，明确对于词汇信息的具体需求，制定科学的审查流程，并及时反馈结果。3）产品架构方面，可尝试加强与用户的交互，精准触达用户需求，提供更加个性化、精细化的服务，比如可提供用户自主设置查阅范围的检索功能，在词库管理界面调整各种词库在词条结构中的显示顺序，还可在词汇查询功能的基础上加入学习板块，帮助用户通过词法、搭配等练习深度理解词条信息、搭建结构化的知识体系。4）在后期维护上，应积极收集用户的使用反馈，不断完善词条内容、提升服务品质，在这个过程中需要和用户建立良好的沟通。

参考文献

[1]Белоусова А. С. Толковые словари [A]. Гл. ред. Ю. Н. Караулов. Русский язык Энциклопедия[Z]. Москва: Большая Российская энциклопедия: Дрофа，1998.

[2]Биржакова Е. Э. Словарь Академии Российской (1789—1794 гг.) [A]. Под ред. Ф. П. Сороколетова. История русской лексикографии[C]. Санкт-Петербург: Наука, 2001.

[3]Лутовинова И. С. Толковые словари[A]. Под ред. Д. М. Поцепни. Лексикография русского языка[C]. Санкт-Петербург: Факультет филологии и искусств СПбГУ, 2009.

[4]Морковкин В. В. Учебники и словари в системе средств обучения русскому языку как иностранному: Сб. ст.[C]. Москва: Русский язык, 1986.

[5]亢世勇. 关于汉语融媒体学习词典的思考——以《当代汉语学习词典》为例[J]. 鲁东大学学报（哲学社会科学版），2020(2).

[6]李　斌，刘雪扬. 基于《汉语大词典》的汉语词汇历时演变计量研究[J]. 南京师大学报（社会科学版），2018(5).

[7]王怿旦，张雪梅. 电子词典检索功能分析及其发展构想[J]. 辞书研究，2016(3).

[8]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).

[9]杨　杨. 18 世纪俄罗斯词典编纂史研究[D]. 黑龙江大学硕士学位论文，2021.

[10]雍和明. 交际词典学[M]. 上海：上海外语教育出版社，2003.

[11]章宜华. 论融媒体背景下辞书编纂与出版的创新[J]. 语言战略研究，2019(6).

[12]章宜华. 融媒体视角下多模态词典文本的设计构想[J]. 辞书研究，2021a(2).

[13]章宜华. 融媒体英语学习词典的设计理念与编纂研究[J]. 外语电化教学，2021b(3).

[14]章宜华. 略论融媒体辞书的技术创新和理论方法[J]. 语言文字应用，2022(1).

[15]郑述谱. 认识历史词典的特点——读《现代俄语历史词源词典》[J]. 辞书研究，1997(3).

The Digitization Development of ***Collection of Russian Language Dictionaries from 1700 to 1850*** and Its Role in the Diachronic Study of Vocabulary

Shi Qi

(Russian Language, Literature and Culture Research Center, Heilongjiang University, Harbin 150080, China)

Abstract: This article focuses on the *Collection of Russian Language Dictionaries from 1700 to 1850* developed by the Kazan Federal University in Russia, introduces its processing of entry texts from source dictionaries and the settings of its retrieval function, and discusses the role of this collection in the study of Russian historical lexicology. Under the development trend of integrated media dictionaries in the digital era, suggestions for further improvement of this collection are put forward.

Keywords: *Collection of Russian Language Dictionaries from 1700 to 1850*; digitization; diachronic study of vocabulary

基金项目：本文系国家社科基金项目“俄罗斯词典编纂史研究”（19BYY208）的阶段性成果。

作者简介：石琦（1997—），女，黑龙江省哈尔滨人，黑龙江大学俄罗斯语言文学与文化研究中心在读硕士，研究方向：词典学

收稿日期： 2023-12-12 **[责任编辑：张春新]**