МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Филологический факультет
Кафедра русского языка
Лаборатория общей и сибирской лексикографии
ИНСТИТУТ РУССКОГО ЯЗЫКА ИМ. В.В. ВИНОГРАДОВА РАН
ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА им. А.С. ПУШКИНА
ПОД ЭГИДОЙ РОССИЙСКОГО ОБЩЕСТВА ПРЕПОДАВАТЕЛЕЙ
РУССКОГО ЯЗЫКА И ЛИТЕРАТУРЫ (РОПРЯЛ)

ЛЕКСИКОГРАФИЯ ЦИФРОВОЙ ЭПОХИ

Сборник материалов Международного симпозиума
(24–25 сентября 2021 г.)

Томск
Издательство Томского государственного университета
2021

MINISTRY OF SCIENCE AND HIGHER EDUCATION
OF THE RUSSIAN FEDERATION
NATIONAL RESEARCH TOMSK STATE UNIVERSITY
Faculty of Philology
Department of the Russian language
Laboratory of General and Siberian Lexicography
V.V. VINOGRADOV RUSSIAN LANGUAGE INSTITUTE
OF THE RUSSIAN ACADEMY OF SCIENCES
PUSHKIN STATE RUSSIAN LANGUAGE INSTITUTE
UNDER THE AUSPIANCE OF THE RUSSIAN SOCIETY OF TEACHERS
OF RUSSIAN LANGUAGE AND LITERATURE (ROPRYAL)

LEXICOGRAPHY OF THE DIGITAL AGE

Proceedings of the International Symposium
(September 24-25, 2021)

Tomsk
TSU Press
2021

УДК 81.374
ББК 81.411.2-4
Л43

Лексикография цифровой эпохи: сборник материалов
Л43 Международного симпозиума (24–25 сентября 2021 г.) / отв. ред. Е.А. Юрина, С.С. Земичева. – Томск : Издательство Томского государственного университета, 2021. – 420 с.
ISBN 978-5-907442-19-1

Сборник содержит материалы Международного научного симпозиума «Лексикография цифровой эпохи» (24–25 сентября 2021 г.), посвященного обсуждению теоретических и прикладных задач лексикографии. В докладах участников отражены тенденции современной лексикографии, представлен научно-практический опыт по созданию лингвистических корпусов, словарей различного типа и их использованию в условиях цифровой среды.

Для исследователей русского языка, лексикографов-практиков, преподавателей вузов, учителей русского языка и литературы, студентов и аспирантов гуманитарных специальностей.

УДК 81.374
ББК 81.411.2-4

ISBN 978-5-907442-19-1

UDC 81.374
LBC 81.411.2-4

Lexicography of the digital age: proceedings of the International Symposium (September, 24-25, 2021) / ed. by E.A. Yurina, S.S. Zemicheva. – Tomsk : TSU Press, 2021. – 420 p.

ISBN 978-5-907442-19-1

The book of proceedings presents materials of the International Scientific Symposium "Lexicography of the Digital Age" (September 24-25, 2021), which brought together lexicographers from around the world to discuss problems related to the creation, use and social purpose of dictionaries in the conditions of civilizational changes of the XXI century. The articles highlight the trends of modern lexicography, scientific and practical experience in creating linguistic corpora, dictionaries of various types and their use for solving research and applied problems.

For researchers of the Russian language, lexicographers-practitioners, university teachers, teachers of the Russian language and literature, students and graduate students of humanitarian specialties.

UDC 81.374
LBC 81.411.2-4

ISBN 978-5-907442-19-1

DOI: 10.17223/978-5-907442-19-1-2021-120

С.О. Савчук
Svetlana O. Savchuk

Электронный словарь вариантов на основе корпуса текстов
Electronic dictionary of variants based on the text corpus

Институт русского языка им. В.В. Виноградова РАН, Москва –
V.V. Vinogradov Russian Language Institute, Moscow
savsvetlana@mail.ru

Аннотация. Электронный словарь вариантов создается на основе текстов XVIII – XX вв. в Национальном корпусе русского языка. Для его формирования используется список несловарных словоформ, снабженных гипотетическими леммами. База данных этих словоформ дает ценный материал для анализа орфографических, фонетических и различных видов грамматических вариантов в языке прошлых эпох.
Summary. Electronic dictionary of variants is being created on the basis of the texts of the 18th -20th century in the Russian National Corpus. For its formation the list of the unidentified word-forms supplied with hypothetic lemmas is used. The database of these word-forms gives valuable material for the analysis of the orthographic, phonetic and grammar variation.
Ключевые слова: Национальный корпус русского языка, орфографические и морфологические варианты, аннотация диахронических корпусов
Keywords: The Russian National Corpus, orthographic and morphological variants, annotation of diachronic corpora

Электронный словарь вариантов создается на базе текстов исторических корпусов (XVIII – 1-й пол. XX в.) в составе Национального корпуса русского языка. Название «Электронный словарь вариантов» понимается в двух смыслах: 1) как составная часть электронного грамматического словаря, с помощью которого выполняется автоматическая аннотация текстов, и 2) как самостоятельный лингвистический ресурс, который может пополнить семейство электронных словарей на основе НКРЯ [5].

Как было показано в работах [1–4], тексты XVIII в. отличаются крайне высокой степенью вариативности, которая слабо отражена в существующих словарях и практически не описана в нормативных грамматиках. Значительная степень вариативности, которая также мало представлена в словарях, отмечена и в текстах XIX и 1-й пол. XX в. Так,

например, слово *фейерверк* присутствует в словаре в единственном орфографическом облике, в то время как в текстах оно может быть передано самыми разными способами: *фейэрверк, фейерверок, феиерверок, фейверк, феерверк, ферверк* и др., ни один из которых не опознается как орфографический вариант одной и той же лексемы.

Все это затрудняет корректную работу морфологического анализатора, поскольку в словаре анализатора вариативность также не учтена. Один из путей формирования словаря морфологического анализатора для диахронического модуля НКРЯ состоит в пополнении словника словами, извлеченными непосредственно из корпуса текстов XVIII–XX в., которые не отражены в лингвистических источниках. Работа в рамках этого направления проводится поэтапно. Основным материалом служат частотные списки словоформ, снабженных их грамматическими характеристиками, полученные на основе обработки текстовых массивов XVIII, XIX и 1-й пол. XX в.

На первом этапе строится частотный словарь словоформ, не получивших предсказанного разбора, которые снабжены гипотетическими разборами.

На втором этапе организуется база данных несловарных словоформ со следующими полями: словоформа, сгенерированная лемма, отсылочная (нормализованная) лемма, грамматические признаки леммы, грамматические признаки словоформы, тип варианта, сведения об авторе и дате создания текста, в котором зафиксирована форма.

Далее проводится анализ и редактирование базы данных: для каждой словоформы выбирается из списка правильная лемма, приписывается отсылочная (нормализованная) словоформа и нормализованная лемма, грамматическая информация проверяется по корпусу, определяются правильные грамматические характеристики.

Следующий этап работы связан с формированием базы данных вариантов. На основании анализа текстов проводится разметка вариантов, выделяются типы вариантов, не учтённых в грамматическом словаре НКРЯ. Среди них самую заметную часть составляют орфографические варианты (*баттарея, галиффэ, забрежжил, всё-же, мятель*), которые бывает трудно характеризовать отдельно от фонетических (*Валенция, гидальго, жужитзу, Мадам-Бутерфлей, музик-холль* и под.). Представлены морфологические (*с одеколонью, к виолончелю, в Марсели,* вм. *с одеколоном*, *к виолончели*, *в Марселе*, *дворяна* вм. *дворяне*, *тетеревей* вм. *тетеревов* и др.) и словообразовательные варианты (*слюдо-*

вой/слюдовый, *степовой, левость, правость, праздношатайство, шапко-закидайство* и под.).

Часть единиц базы данных отбирается для пополнения словаря морфологического анализатора. Основными критериями отбора являются частотность словоформы, формальный облик слова (если внешний облик словоформы не дает возможности точно предсказать его лемму и грамматические характеристики, слово включается в словарь), наличие вариантов также является основанием для помещения слова в словарь.

Электронный словарь вариантов, созданный на основе НКРЯ, будет полезен для изучения особенностей языка предшествующих эпох и истории становления литературной нормы.

Литература

1. *Поляков А.Е., Савчук С.О., Сичинава Д.В.* Грамматический словарь для автоматического анализа текстов XVIII–XIX века: первые результаты // Компьютерная лингвистика и интеллектуальные технологии. Вып. 12 (19). М., 2013. С. 574–586.

2. *Савчук С.О., Сичинава Д.В.* Корпус русских текстов XVIII века в составе НКРЯ: проблемы и перспективы // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 52–70.

3. *Савчук С.О.* Электронный словарь вариантов на основе текстов XVIII века // Информационные технологии и письменное наследие: материалы IV международной научной конференции (Петрозаводск, 3–8 сентября 2012 г.). Петрозаводск; Ижевск, 2012. С. 241–244.

4. *Савчук С.О.* Из опыта разработки автоматического морфологического анализатора для текстов XVIII-XIX века // Писменото наследство и информационните технологии Текст: материали от V международна науч. конф. (Варна, 15–20 септември 2014 г. / отг. ред. В.А. Баранов, В. Желязкова, А.М. Лаврентьев. София; Ижевск, 2014. С. 138–142.

5. Словари, созданные на основе Национального корпуса русского языка. URL: http://dict.ruslang.ru (дата обращения: 20.01.2021).