УДК 81’33

Горбань Оксана Анатольевна
доктор филологических наук,
профессор кафедры русской филологии
и журналистики
Волгоградский государственный
университет
Волгоград, Россия

Oksana Gorban
Doctor of Philology
Professor of Department
of Russian Philology and Journalism
Volgograd State University
Volgograd, Russia
oa_gorban@volsu.ru

Косова Марина Владимировна
доктор филологических наук,
профессор кафедры русской филологии
и журналистики
Волгоградский государственный
университет
Волгоград, Россия

Marina Kosova
Doctor of Philology
Professor of Department
of Russian Philology and Journalism
Volgograd State University
Volgograd, Russia
mv_kosova@volsu.ru

Шептухина Елена Михайловна
доктор филологических наук,
профессор кафедры русской филологии
и журналистики
Волгоградский государственный
университет
Волгоград, Россия

Elena Sheptukhina
Doctor of Philology
Professor of Department
of Russian Philology and Journalism
Volgograd State University
Volgograd, Russia
em_sheptuhina@volsu.ru

ЛИНГВИСТИЧЕСКИЕ ПРОБЛЕМЫ СОЗДАНИЯ КОРПУСА ДОКУМЕНТОВ ВОЙСКА ДОНСКОГО (XVIII–XIX ВВ.)

В статье рассматриваются подходы к решению таких лингвистических проблем создания диахронического корпуса документов канцелярий Войска Донского из Государственного архива Волгоградской области, как отбор источников, принципы передачи текстов с устаревшей графикой, параметры аннотирования и структурной разметки. Обоснована релевантность включения в корпус всех видов документов разного объема и структуры текста, необходимость частичной адаптации графики с сохранением вышедших из употребления букв, орфографии и пунктуации оригиналов как значимых для исторического языкознания, исторической диалектологии русского языка. На основе систематизации стандартизированных речевых оборотов, эксплицирующих элементы аннотирования, показаны возможности автоматизации процесса метаразметки. При отсутствии в текстах последовательной постановки знаков препинания, а также с учетом особенностей синтаксиса и композиции документов предложено при структурной разметке деление текстов на структурно-смысловые блоки, которые могут быть расчленены на менее объемные синтаксические единицы.

К л ю ч е в ы е с л о в а: *русский язык, лингвистический корпус, диахронический корпус, архивные документы, адаптация текстов, метаразметка, структурная разметка.*

LINGUISTIC ISSUES OF CREATING A CORPUS OF THE DON COSSACK ARMY DOCUMENTS (18th–19th CENTURIES)

The article discusses approaches to solving such linguistic problems of creating a diachronic corpus of documents of the Don Cossack Army offices from the State Archive of the Volgograd region, as the selection of sources, principles of transmission of texts with outdated graphics, annotation parameters and structural markup. The relevance of including all types of documents of different text volumes and structures in the corpus, the need for partial adaptation of graphics while preserving obsolete letters, spelling and punctuation of the originals as significant for historical linguistics, historical dialectology of the Russian language is substantiated. Based on the systematization of standardized speech phrases that explicate annotation elements, the possibilities of automating the meta-tagging process are shown. In the absence of consistent punctuation marks in the texts, as well as taking into account the peculiarities of syntax, the composition of documents is proposed to divide texts into structural and semantic blocks, which can be divided into less voluminous syntactic units.

K e y w o r d s: *Russian language; linguistic corpus; diachronic corpus; archival documents; text adaptation; meta-markup; structural markup.*

Деловые документы XVIII века представляют ценность в качестве источников изучения начального этапа формирования национального литературного языка. Именно в официально-деловой письменности, как отмечают исследователи, складывались и закреплялись новые литературно-языковые нормы [1, с. 72; 2, с. 156–157]; здесь в первую очередь реализовалась языковая политика петровской эпохи, проявлявшаяся в экспансии заимствований, конкуренции заимствованной и исконной управленческой терминологии [3, с. 984–989; 4] и др. Раскрыть во всей полноте многообразие функционирования русского языка с учетом как общерусских, так и локальных речевых традиций и тенденций языкового развития позволяет обращение к источникам, создававшимся в различных регионах, в частности

к документам учреждений Войска Донского. При изучении истории Войска возникает целый ряд источниковедческих проблем, одна из которых – утрата значительной части войскового архива (см.: [5]). В связи с этим особую значимость приобретает публикация региональных деловых письменных памятников, а также создание электронных корпусов документов, в том числе корпусов лингвистических.

Документы Войска Донского XVIII–XIX веков вызывают интерес лингвистов по ряду причин. После того, как в начале XVIII века регулирование отношений России с Донским казачьим войском было передано из Коллегии иностранных дел в Военную коллегию, Войско, по сути, утратило государственную автономию и вошло в Российскую империю с сохранением некоторое время автономии областной. Изменились структура управления, порядок подчинения, обязанности казаков и т. п. [6, с. 220–227]. Это повлияло на делопроизводство Донского казачьего войска, язык и стиль войсковых документов, в которых проявились общие тенденции литературного языка и особенности живой речи казаков.

Создание лингвистического корпуса этих источников открывает дополнительные возможности для исследований в области истории русского языка вообще и южнорусских говоров Нижней Волги и Дона в частности. Такая работа ведется в Волгоградском государственном университете.

Корпус включает документы архивного фонда «Михайловский станичный атаман» Государственного архива Волгоградской области (ГАВО, фонд 332, опись 1) – источники 1734–1837 гг., созданные в войсковой и станичных канцеляриях, некоторых учреждениях других регионов. Это не публиковавшиеся ранее документы разных жанров, написанные скорописью первой и второй половины XVIII в. и первой половины XIX в. Общий объем составляет ок. 10200 рукописных листов, по предварительной оценке, около 10,3 млн словоформ, включая служебные слова.

Исследования документов архивного фонда были поддержаны совместными грантами Российского гуманитарного научного фонда и Администрации Волгоградской области в 2013–2014 (№ 13-14-34008), 2016–2017 (№ 6-14-34004) гг. и проводились коллективом в составе: О. А. Горбань, М. В. Косова, Е. М. Шептухина, И. С. Герасимова, на разных этапах также Е. Г. Дмитриева, И. А. Сафонова, Э. У. Саидгасанова, Е. Л. Берестова. Привлекались студенты отделений филологии и прикладной математики Е. С. Балясова, Д. С. Бондарева, А. С. Комендантов. Результатом стала коллективная монография, включающая опубликованные документы (дела 1–9 из 154) и их лингвистическое описание в аспекте жанровых параметров, использования различных лексических единиц, стилистических средств и т. д. [7].

В 2019–2021 гг. проект, связанный с созданием лингвистического корпуса архивных документов, был поддержан Российским фондом фундаментальных исследований (№ 19-012-00246). Участники проекта: О. А. Горбань,

М. В. Косова, Е. М. Шептухина, А. В. Светлов, И. С. Герасимова, Н. И. Тихонова, с привлечением студентов отделения прикладной математики А. Г. Матвеева, Д. Ю. Филимонова, Ю. Д. Сапич, А. В. Павлова.

В ходе работы потребовалось решение ряда задач лингвистического характера: 1) сформулировать принципы отбора источников для корпуса; 2) подготовить тексты для машинной обработки; 3) определить релевантные для корпуса параметры метаразметки текстов и выявить возможности ее автоматизации; 4) произвести структурную разметку. В соответствии с принятыми решениями специалистами в области IT-технологий созданы новые или адаптированы существующие программные продукты.

1. Принципы отбора источников для корпуса

В качестве источников для лингвистического корпуса наиболее ценными представляются документы, содержащие развернутый, связный текст, который раскрывает семантику и функционирование языковых единиц. Таковыми являются войсковые грамоты, паспорта (пашпорты), сказки и другие входящие и исходящие документы, обеспечивающие коммуникацию между учреждениями, должностными и иными лицами. Однако полнота отражения в корпусе русского языка донского региона XVIII–XIX вв. может быть достигнута только за счет привлечения всех жанров документов, в том числе и таких, которые содержат лишь отдельные связные предложения, а преимущественно – перечни имен лиц, наименований предметов. Это списки казаков, ведомости прихода-расхода денег, расписки о получении товара, денег за товары. Так, списки (реестры) казаков предоставляют обширный материал для исторической ономастики (антропонимики), включают специфическую терминологию, связанную с особенностями структуры войска, например: *сказочные казаки* – приписные, осевшие на Дону пришлые люди, не входящие в списочный состав Войска, *действительные казаки* – служилые казаки, входящие в так называемые списки, в основной состав Войска, *малолетки*, *выростки*, *невыростки* – разные категории молодых казаков, еще не давших присягу (*невыросток* не встретилось в известных нам лексикографических источниках) и др. Различные учетные документы (расписки, ведомости и под.) часто содержат лексические единицы, принадлежащие разным языковым пластам – книжному, разговорному, включающему и диалектную лексику.

При передаче текстов и их последующей разметке сохраняются также разного рода делопроизводственные записи на полях, поскольку они могут включать термины, в том числе редкие, как, например, *трибликатная*.

В фонде хранятся не только беловики, но и черновики документов. Так, документы, исходящие из канцелярии Михайловской станицы, отложились в фонде преимущественно в виде не отпусков или копий, а черновиков. В связи с этим стоят вопросы о включении их в корпус и о форме представления в нем. Обнаружено, что зачеркнутые фрагменты могут содержать редкую диалектную лексику, отсутствующую в окончательном варианте документа после его редактирования, которая имеет ценность для диалектологов,

историков языка. На данный момент решено черновики использовать в частично восстановленном виде: вставки включить в основной текст, не выделяя их особо, а зачеркнутые фрагменты сохранять, учитывать при морфологической разметке и представлять их в результатах поискового запроса как зачеркнутые. Например, в одном из доношений станичного старшины употреблено диалектное слово *бахча*, затем оно зачеркнуто и сверху написано общерусское *огород*. При включении в корпус только белового текста зачеркнутое *бахча* будет утрачено; предполагаемый нами вариант подачи черновика позволит его отразить в корпусе: *будучи он Михеевъ у меня Лащилина ~~на бахче~~ на ωгороде…* (ГАВО, ф. 332, оп. 1, д. 8, л. 4 об.). Технически эта задача пока не решена.

2. Подготовка текстов документов для машинной обработки

Как отмечалось выше, документы написаны скорописью XVIII века. В связи с этим при переводе текстов в машиночитаемый формат проведена их адаптация, первый этап которой заключается в раскрытии тител, восстановлении утраченных фрагментов, раздельном написании предлогов и частиц, написании имен собственных с прописной буквы и некоторые другие. Следующий этап адаптации должен заключаться в выборе графической системы и орфографических принципов передачи текстов. Как известно, в Национальном корпусе русского языка все тексты XVIII и XIX вв. передаются средствами современной орфографии, поскольку, как подчеркивают исследователи, «она лежит в основе всех программных инструментов разметки текстов» [8, с. 57]. В создаваемом корпусе войсковых документов, кроме указанных выше изменений, передается графика и орфография оригиналов: сохранены буквы ѕ, ѡ, ѣ, і, ѳ, ѯ, написания, отражающие живое произношение, и др. Это, по мнению коллектива, позволит использовать корпус как источник изучения письменной традиции Юга России, особенностей живой речи, варианты, свидетельствующие о становлении орфографической нормы.

Например, фрагмент оригинального скорописного документа при точной передаче в компьютерном наборе выглядит следующим образом:

Благоро^{д}ны ипочтѣнны г$^{с\tilde{}}$днъ капита^{н}
В хоперскои крѣпости камендантъ
анамъ милостивы
г$^{с\tilde{}}$дрь иванъ
андрѣевичъ
Писмо о^{т}вашего благородия по^{лу}чилъ
вкоторо^{м} написано …
(ГАВО, ф. 332, оп. 1, д. 7, л. 5)

В корпусе текст представлен в таком виде:
Благородны и почтѣнны г(о)с(по)д(и)нъ капитан в хоперскои крѣпости камендантъ а намъ милостивы г(о)с(у)д(а)рь Иванъ Андрѣевичъ
Писмо от вашего благородия получилъ в котором написано…

Для того, чтобы обеспечить морфологическую разметку текстов с устаревшей графикой, IT-специалистами ВолГУ создано приложение к утилите MyStem И. Сегаловича, которое описано в [9]. Использование орфографии оригиналов потребовало при раскрытии тител восстанавливаемые буквы заключать в скобки с целью разграничения оригинальных и реконструируемых написаний. Корректное отражение при разметке слов с внутренними скобками на сегодня ждет своего программного решения.

Тексты документов, особенно середины XVIII века, содержат устаревшие грамматические формы, которые могут быть определены программой ошибочно. Например: дат. пад. мн. числа с окончанием *-ом* (вместо современного *-ам*); творит. пад. мн. числа с окончанием *-ы* (вместо современного *-ами*), форма аориста *умре* и нек. др. При отсутствии единых орфографических правил наблюдаются вариантные написания окончаний род. пад. ед. числа прилагательных *-ого*/*-аго*/*-ова*, при этом формы на *-ова* могут определяться как формы фамилий на *–ов* и как формы прилагательных (например, *Толстова* может быть соотнесено с леммой *Толстов* и *Толстой*). Все это осложняет морфологическую разметку текстов. Созданное приложение имеет функционал для снятия омонимии вручную, если автоматическими средствами морфологические характеристики слова определены неверно. Однако в зависимости от частотности названных единиц решается, достаточно правки вручную или потребуется корректировка программы.

3. Параметры аннотирования (метаразметки)

Документный текст имеет характерную структурно-композиционную и речевую организацию, детерминированную не только его общими свойствами, но и требованиями жанра, – формуляр, типовые речевые формулы и т. д. С опорой на предложенную Т. В. Шмелевой [10] модель речевого жанра членами коллектива разработана жанровая модель документного текста, которая позволяет с единых позиций выявить обязательные и специфические черты документа, установить его жанр. Модель включает такие жанровые параметры, как «адресант», «адресат», «функция», «характер передаваемой информации», «структура», «доминирующая модальность», «пространственная локализация документа (пространство)», «временная локализация документа (время)» [11; 12]. Эти параметры соотносятся, с одной стороны, с текстовыми категориями, с другой – с элементами формуляра (реквизитами) документа. В метаразметку включены параметры «адресант», «адресат», «место» и «дата создания», «место» и «дата получения» документа. Указывается также жанр (вид) документа, есть или отсутствует печать, является ли документ подлинником, копией, черновиком и т. д., место хранения (архивный адрес).

Указанные реквизиты, как правило, выражаются устойчивыми словосочетаниями, клишированными оборотами, которые способствуют стандартизации текста документа и служат маркерами жанра. В XVIII–XIX веках такая стандартизованность была свойственна многим видам документов. Это позволило предположить, что выявление типовой структуры текстов и ее вербальных маркеров может быть использовано для автоматизации процесса

метаразметки документов. Например, в тексте рапорта выделяются следующие формулы: *покорнеиши* ***репортъ*** (жанр документа), *Воиска Данскаго* ***воисковому атаману*** *... и всему* ***Воиску Донскому*** (адресат), *о семъ покорнеиши репортуют* <имя и фамилия в им. пад.> (адресант), <число> *году* <название месяца> <число> *дня* (дата) и др.

Главный идентификатор жанра (вида) документа – его название, в тексте это самоназвание, однако не все документы его содержат. Проблема жанровой идентификации заключается в отсутствии не только самоназвания в текстах, но четких жанровых границ, однозначных определений видов документов в рассматриваемый исторический период. Наблюдается вариативность документов одного жанра, совпадение функций разных документов (список, опись), наличие синонимичных названий (старых и новых – *доношение* и *репорт*, *список* и *реестр* и др.). Это требует дополнительной исследовательской работы. На настоящем же этапе по результатам анализа текстов систематизированы речевые маркеры выбранных для аннотирования параметров, составлены обобщающие таблицы, на основе которых IT-специалистами предпринята попытка автоматической идентификации вида документа, адресата, адресанта и других метаданных. Создано приложение, позволяющее по обнаруженным шаблонам определить жанр и другие параметры аннотирования; оно описано в [13]. Корректная работа приложения достигнута для войсковой грамоты, рапорта, доношения, известия.

6. Структурная разметка текстов

Синтаксис документов носит книжный характер. Многие предложения осложнены однородными членами, уточняющими оборотами и другими конструкциями. Часто используются сложные предложения с разными видами синтаксической связи, сложноподчиненные предложения с несколькими придаточными. В текстах середины XVIII века сохраняется такая старая особенность синтаксиса, как нанизывание предложений при помощи начального союза *а*. Знаки препинания функционально слабо дифференцированы, границы предложений часто не маркированы, а постановка точки не всегда совпадает с концом предложения. Все это затрудняет синтаксическое членение текста и требует специального исследования синтаксических особенностей документов. На данном этапе работы в качестве единицы текста принимается структурно-смысловой блок, внутри которого в ряде случаев возможно выделение менее объемных синтаксических единиц [14]. Приведем фрагмент войсковой грамоты, где знаком **&** показано предлагаемое деление текста на сегменты:

Сего ѳевраля 1 дня написанноі в Верхнюю Рыковскую станіцу в число казаковъ Спиридонъ Ларионовъ с(ы)нъ Матаригинъ о подлиннои ево родине в канцелярііі воисковыхъ делъ допросом показалъ **&** родился де онъ на Хопре в Михаиловскои станицы казачеи с(ы)нъ ıs вашеи де Михаиловскои станицы в прошломъ 1745ом году еще приживности о(т)ца ево Лариона Никиѳорова /: **&** а ныне какъ онъ слышел что онои уже умре :/ съехалъ онъ по бѣдности ихъ вашеи же Михаиловскои станицы с казакомъ Аѯеномъ Лащилинымъ в город Черкаскои где от него отставъ жителствовал

по разным людемъ :/ **&** і в прошлом 753м году женился онъ в Нижнеи Рыковскоі станіцы на девке казачеи дочери где иімѣет жителство до сего времѣни (ГАВО, ф. 332, оп. 1, д. 11, л. 2).

Результаты решения ряда обозначенных задач являются основой создаваемого специализированного диахронического лингвистического корпуса документов. Такой корпус может служить источником для исторического языкознания, диалектологии, дипломатики, отечественной истории.

ЛИТЕРАТУРА

1. Виноградов В. В. Очерки по истории русского литературного языка XVII–XIX веков. М. : Высшая школа, 1982. 529 с.
2. Марков В. М. Проблемы грамматической лексикологии и русский литературный язык XVIII века // Избранные работы по русскому языку. Казань, 2001. С. 156–163.
3. Живов В. М. История языка русской письменности : в 2 т. Т. 2. М. : Русский фонд содействия образованию и науке, 2017. 480 с.
4. Шамшин И. В. Иноязычные наименования документов в русской административной лексике XVIII века // Вестник Московского государственного областного университета. Сер. Русская филология. 2007. № 1. С. 154–157.
5. Сень Д. В. Архив Войска Донского и история войскового делопроизводства: актуальные вопросы изучения // Научное наследие профессора А. П. Пронштейна и актуальные проблемы развития исторической науки (к 95-летию со дня рождения выдающегося российского ученого). Ростов-на-Дону, 2014. С. 484–495.
6. Пронштейн А. П. Земля Донская в XVIII веке. Ростов-на-Дону : Изд-во Рост. ун-та, 1961. 375 с.
7. Документы Войска Донского XVIII века: лингвистическое описание и тексты : монография / О. А. Горбань, М. В. Косова, Е. М. Шептухина, Е. Г. Дмитриева, И. А. Сафонова. Волгоград : Изд-во ВолГУ, 2020. 464 с.
8. Савчук С. О., Сичинава Д. В. Корпус русских текстов XVIII века в составе Национального корпуса русского языка: проблемы и перспективы // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб., 2009. С. 52–70.
9. Комендантов А. С., Матвеев А. Г., Светлов А. В. Автоматизация морфологической разметки архивных документов [Электронный ресурс] // Математическая физика и компьютерное моделирование. 2019. Т. 22. № 4. С. 53–63.

DOI: https://doi.org/10.15688/mpcm.jvolsu.2019.4.4

10. Шмелева Т. В. Модель речевого жанра // Жанры речи. Саратов, 1997. Вып. 1. С. 88–98.
11. Жанровые особенности войсковых грамот середины XVIII в. (по материалам архивного фонда «Михайловский станичный атаман») [Электронный ресурс] / О. А. Горбань, Е. Ю. Ильинова, М. В. Косова, Е. М. Шептухина // Известия Уральского федерального университета. Сер. 2. Гуманитарные науки. 2016. Т. 18. № 4 (157). С. 182–199.

DOI: https://doi.org/10.15826/izv2.2016.18.4.074

12. Cossack Military Charters of the mid18th Century: Genre Distinction [Electronic resource] / O. A. Gorban, E. Yu. Ilyinova, M. V. Kosova, E. M. Sheptukhina // XLinguae Journal. 2017. Vol. 10, Issue 3. P. 123–136.

DOI: https://doi.org/10.18355/XL.2017.10.03.10

13. Автоматизация процесса метаразметки архивных документов [Электронный ресурс] / Д. Ю. Филимонов, А. В. Светлов, О. А. Горбань, М. В. Косова, Е. М. Шептухина // Математическая физика и компьютерное моделирование. 2020. Т. 23, № 4. С. 57–69.

DOI: https://doi.org/10.15688/mpcm.jvolsu.2020.4.6

14. Горбань О. А., Косова М. В., Шептухина Е. М. Структурная разметка деловых документов в диахроническом лингвистическом корпусе: проблемы и решения [Электронный ресурс] // Вестник Волгоградского государственного университета. Сер. 2, Языкознание. 2021. Т. 20, № 4. С. 5–18.

DOI: https://doi.org/10.15688/jvolsu2.2021.4.1