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

УДК 004.91, 81'33, 004.42
ББК 32.973, 81.1

Дата поступления статьи: 02.09.2020
Дата принятия статьи: 17.11.2020

АВТОМАТИЗАЦИЯ ПРОЦЕССА МЕТАРАЗМЕТКИ АРХИВНЫХ ДОКУМЕНТОВ [1]

Даниил Юрьевич Филимонов

Студент института математики и информационных технологий,
Волгоградский государственный университет
dane020597@mail.ru, matf@volsu.ru
просп. Университетский, 100, 400062 г. Волгоград, Российская Федерация

Андрей Владимирович Светлов

Кандидат физико-математических наук,
доцент кафедры математического анализа и теории функций,
Волгоградский государственный университет
a.v.svetlov@gmail.com, andrew.svetlov@volsu.ru, matf@volsu.ru
https://orcid.org/0000-0002-8764-6132
просп. Университетский, 100, 400062 г. Волгоград, Российская Федерация

Оксана Анатольевна Горбань

Доктор филологических наук,
профессор кафедры русской филологии и журналистики,
Волгоградский государственный университет
oa_gorban@volsu.ru
https://orcid.org/0000-0002-2345-3673
просп. Университетский, 100, 400062 г. Волгоград, Российская Федерация

Марина Владимировна Косова

Доктор филологических наук,
профессор кафедры русской филологии и журналистики,
Волгоградский государственный университет
mv_kosova@volsu.ru
https://orcid.org/0000-0003-2854-8759
просп. Университетский, 100, 400062 г. Волгоград, Российская Федерация

Елена Михайловна Шептухина

Доктор филологических наук,
профессор кафедры русской филологии и журналистики,
Волгоградский государственный университет
em_sheptuhina@volsu.ru
https://orcid.org/0000-0002-8007-6042
просп. Университетский, 100, 400062 г. Волгоград, Российская Федерация

Аннотация. Работа посвящена описанию созданного авторами статьи приложения для поиска определенных заранее заданных шаблонов в обрабатываемых текстах. Данные шаблоны описываются специально разработанной системой маркеров, специфических для ряда рассматриваемых документов. Основное назначение данного приложения — подготовка метаразметки документов архивного фонда «Михайловский станичный атаман» для создания лингвистического корпуса. В ходе работы над приложением была решена задача корректного определения документов четырех типов — войсковая грамота, рапорт, доношение и известие — а также их описательных характеристик.

Ключевые слова: автоматизация лингвистического анализа, автоматизация метаразметки, графический интерфейс, корпусная лингвистика, регулярные выражения.

Введение

Задачи анализа текстов, возникающие в лингвистике, часто требуют выполнения довольно большого объема рутинных операций, которые успешно поддаются автоматизации посредством специализированных компьютерных программ. Причем в настоящее время круг задач, решение которых можно передать ЭВМ, включает не только относительно простые, типа построения частотного анализа или синтаксического и морфологического разбора текста, но и более сложные, такие как семантический анализ, автоматическое определение стиля текста или даже его возможного автора. Препятствием на пути автоматизации обработки может быть, пожалуй, только недоступность электронной формы документа.

Такие документы, содержащие ценные для истории русской культуры и русского языка сведения, в основном хранятся в архивах и не используются широкой научной общественностью из-за их малой известности. Таким примером являются документы фонда «Михайловский станичный атаман» (1734–1836 гг.), хранящиеся в Государственном архиве Волгоградской области (ГАВО. Ф. 332. Оп. 1). Данные документы были написаны с использованием скорописи первой и второй половины XVIII в., для которой не существует компьютерных средств для распознавания. Коллектив ученых ВолГУ под руководством профессора О.А. Горбань провел транслитерацию текстов, при этом была сохранена орфография оригиналов: соблюдены выносные буквы строка в строку (буквы, которые пишутся над строкой, хотя и составляют часть слова, остающегося в строке), сохранено титло (надстрочный знак над сокращенно написанным словом), другие надстрочные знаки, все пометки на полях были даны отдельными строками с соответствующим комментарием в сносках (см.: [11; 12]). Следующим этапом на пути

к автоматизированному анализу в аспекте корпусной лингвистики стала адаптация текстов: выносные буквы даны в строку, титла раскрыты, диграфы устранены, добавлены пробелы после предлогов и перед энклитическими частицами. В нынешнем виде эти документы стали пригодны для компьютерной обработки без необходимости создания специализированных средств, корректно работающих, например, с выносными буквами и титлами: можно использовать стандартные инструменты для обработки текстов, создавая новое программное обеспечение для решения специфических задач, возникающих при подготовке документов для включения их в языковой корпус. Одной из таких задач и посвящена данная работа.

1. Постановка задачи

Как уже было замечено выше, основной целью данного проекта является создание корпуса документов фонда «Михайловский станичный атаман». Методы корпусной лингвистики представляются наиболее оптимальными в данном случае, так как предполагают обработку большого количества текстов с целью решения самых разнообразных лингвистических задач. Наша группа присоединилась к коллективу филологов для обеспечения технической и программной части проекта. Главной задачей для нас является создание «движка» корпуса документов, то есть программного обеспечения, решающего задачи хранения базы данных размеченных текстов, выполнения запросов к этой базе, а также предоставления пользователям удобного интерфейса для работы, не требующего специальной квалификации в области информационных технологий. Однако параллельно с созданием «движка» необходимо провести подготовку документов к включению в корпус: все тексты должны пройти специальную обработку — разметку.

Существует большое число разновидностей разметки, и в предыдущих публикациях [6; 9] нашей группы уже было описано решение задачи морфологической разметки. Настоящая статья посвящена вопросу метаразметки. Под метаразметкой понимается приписывание тексту определенных описательных атрибутов. В случае делопроизводственных документов это в первую очередь такие параметры, как тип документа (жанр), автор (составитель, адресант), адресат, дата и место создания. Метаразметка необходима прежде всего для реализации поискового аппарата корпуса, чтобы исследователь, пользующийся им, мог получать выборки текстов с заданными внешними параметрами: например, тексты определенного типа, созданные в определенный период, направленные определенному адресату и т. п.

Решению задачи автоматического определения жанра и автора текста посвящено довольно много исследований, в результате которых разработано большое количество разнообразных методов на основе статистики (кластерного и дискриминантного анализа), нейронных сетей, а иногда и их сочетаний [1; 2; 4; 5; 8; 13; 15].

Однако перед нами стоит гораздо более простая задача: в архиве фонда «Михайловский станичный атаман» хранятся в основном документы канцелярий Войска Донского середины XVIII – первой трети XIX в., и разновидностей этих документов не так много. Более того, поскольку это чаще всего официальные документы, составлены они были по определенным шаблонам, формам, параметры которых сравнительно легко можно извлечь из документов посредством предварительного анализа. Эта работа также проводится коллективом ученых-филологов ВолГУ под руководством профессора О.А. Горбань. Результатом проведенной ими систематизации документов стало выделение специальных речевых маркеров жанровых параметров всех встречающихся в архиве

типов документов [3; 7; 10; 14]. На рисунках 1–2 представлены фрагменты описания маркеров для документов некоторых типов.

Параметры	Значения параметров и текстовые маркеры
Вид/жанр	Войсковая грамота [в тексте]: словосочетание *сия войсковая грамота* или *сия грамота* в любом падеже: *- сия н(а)ша* ***воисковая грамота*** *- с сеи н(а)шаї* ***грамоты*** [конечная фраза]: ***у сеи грамоты*** *н(а)ша Воиска Донскаго печать*
Адресант	Войско Донское [начальная фраза] *Ѿт донских атаманов и казаков от (...) воискового ... атамана ... от всего* ***Воиска Донскаго*** [в тексте]: *мы* + ***Войско Донское***: [конечная фраза]: *у сеи грамоты н(а)ша* ***Воиска Донскаго*** *печать*
Адресат	[продолжение начальной фразы]: Дат. падеж *атаману (атамана) и казакам* (+ названия станицы или станиц как уточнение адресата); Дат. падеж названия должности + (имя): *- Михайловской станицы станичному* ***атаману и казакам****;* *- по Хопру от Букановской до Михайловской станицы станичнымъ* ***атаманомъ и казакамъ****;* *- упровляющему почьтъместерскую должность казаку* ***Сеновию Рвачеву***
Место создания	Черкасскиї [конечная фраза]: *писана в* ***Черкаскомъ***

Рис. 1. Фрагмент описания маркеров для документов типа «Войсковая грамота»

Параметры	Значения параметров и текстовые маркеры
Вид / жанр	Известие [в начальной фразе после указания адресанта (*от* + Род. падеж) и адресата (*к* + Дат. падеж) заголовок]: *- от ... старшины Петра Лащилина ... к военному суду* ***известие****;* [в середине текста]: словосочетание *сие известие* в любом падеже [в конце текста перед подписью]: *сим известиемъ + сообщается*: *-и о том оному военному суду сим* ***известиемъ*** *от меня сообщяетца*
Адресант	[начальная фраза]: *от* + Род. падеж (чин + имя + фамилия): *- от ωпределеннаго от Воиска Донскагω депутатом* ***старшины Петра Лащилина****;* [конечная фраза при подписи]: Им. падеж (чин + имя) или *вместо онаго* + Род. падеж (чин, фамилия): *- Вместо онаго* ***старшины Лащилина*** *за неумением ево грамоты подписал находящеися при нем козакъ Евимъ Немухин*
Адресат	[начальная фраза]: *к* + Дат. падеж (чин и имя лица или название учреждения); *-* ***къ военному суду*** *известие;* [в конце текста]: Дат. падеж (чин и имя лица или название учреждения) перед словами *сим известием от меня сообщается:* *- и о том оному* ***военному суду*** *симъ известием от меня сообщяетца*

Рис. 2. Фрагмент описания маркеров для документов типа «Известие»

Непосредственно пример войсковой грамоты с отмеченными маркерами приведен на рисунке 3.

л. 5	
Ѿ*т* донских атаманов и казаков ***от*** войскового	адресант
опред*ѣ*ленного до указу атамана Ивана Ивановича	
с(ы)на Ѳролова и **от** всего **Войска Донскаго**: по Хопру	
от **Букановскоi до Михаиловскоi станицы станищным**	
атаманомъ і казакамъ об*ъ*являем. сего дека-	адресат
бря 9*г(о)* дня по нашему войскового отамана	
разсмотрению предъявленѡ было мною жъ ата-	
маномъ при собранiї войскового круга всем*у*	
Донскому Войску что н(ы)нешней под Крым поход	
с атаманом Иваномъ Ханжонкомъ казакам	
кои в томъ походе были за службу *л(ъ)* или по-	
ход причтетца; и на то требованѡ	
определения дабы в Войске Донском впред	
в нарядах затруднения: и ко ѡтправлению	
по указомъ в службы і в походы остонѡ-	
вки быть не могло, того ради приговорили	
мы **Войском Донскимъ** в своемъ войсковомъ	адресант
кругу оное камандрование с походным ата-	
маномъ Іваном Ханжонком всемъ казакам	
кои в томъ камандрованiї были зачитат*(ъ)*	
за поход: того *же* ради для ведома и испол-	
нѣния в каждоi станицы описыват*(ъ)* с сеи	
н(а)шаї **грамоты** кои потомъ исполнять	вид документа
непременно а по другим рекам такѡжъ	
л. 5 об.	
і по дешним станицам к атаманом і каза-	
кам ко исполнению того жъ н(а)ши войсковыя	
грамоты и приказы посланы и какъ вы	
которою станицею сию н(а)шу **войсковую гра-**	
моту получите і вамъ чинить по выше-	вид документа
писанному непременно писана в **Черка-**	
скомъ 1735г(о) году декабря 17г(о) дня. У сей	место и дата
грамоты наша **Войска Донскаго** печать	адресант

Рис. 3. Пример войсковой грамоты с разметкой

Таким образом, оказалось, что в нашем случае нет необходимости в трудоемких методах статистического анализа или машинного обучения, достаточно провести в документе поиск определенных маркеров. Причем главным маркером во всех рассмотренных документах является прямое указание их вида. Прочие маркеры являются вспомогательными элементами метаразметки. Однако стоит заметить, что среди этих маркеров есть как универсальные, не зависящие от вида документа (например, дата создания), так и специфические, изменяющие свою форму в документах разных жанров (например, адресант в войсковых грамотах — всегда Войско Донское, а в документах других типов могут встречаться разные адресанты, и способ их указания в документах также отличается). Поэтому алгоритм получения метаразметки должен выглядеть как представлено на рисунке 4.

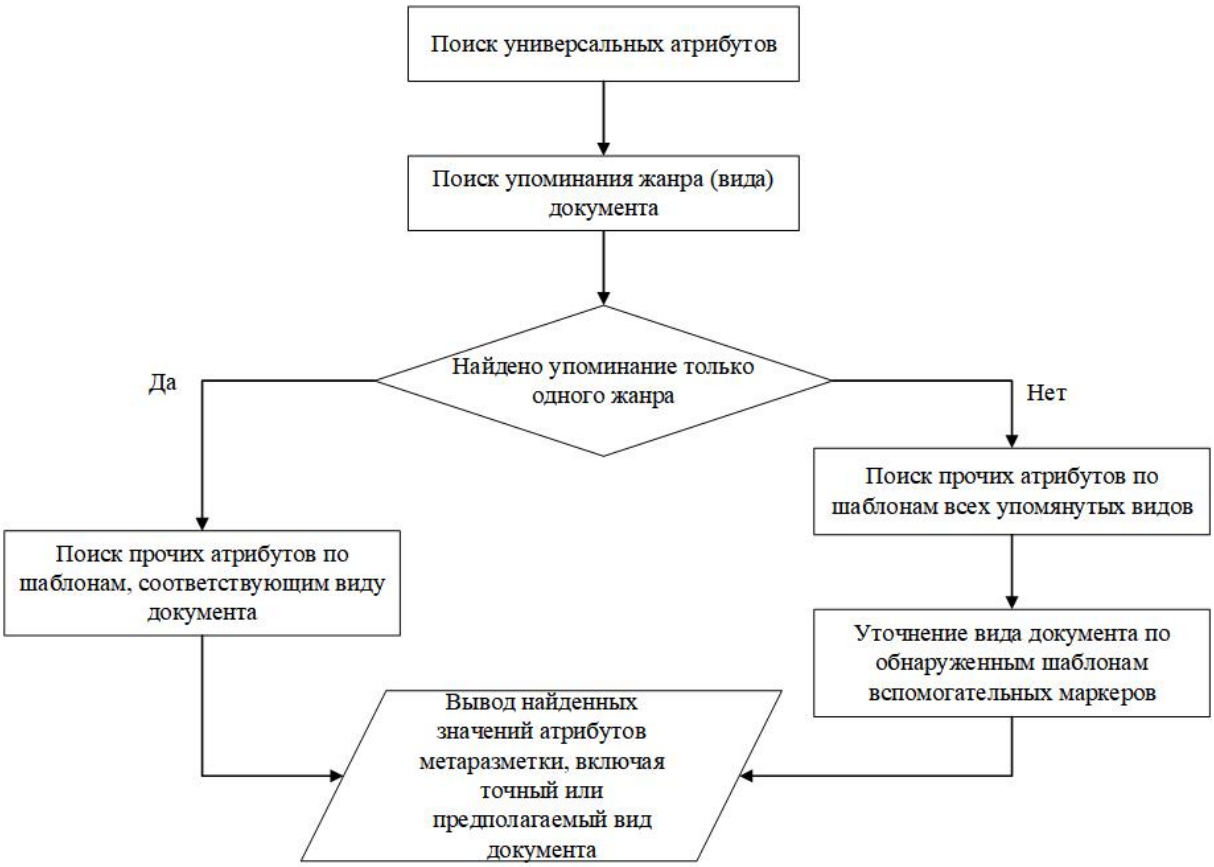

Рис. 4. Алгоритм извлечения метаразметки

2. Разработка регулярных выражений

Итак, как уже было отмечено выше, в качестве исходных материалов для создания приложения нами были получены таблицы маркеров, встречающихся в документах разных типов. На данный момент в нашем распоряжении имеются параметры метаразметки следующих жанров: «войсковая грамота», «рапорт», «доношение» и «известие». В соответствии с представленным алгоритмом (см. рис. 4) основной функцией программы будет поиск маркеров разных видов. Очевидным решением этой задачи является использование регулярных выражений.

Для составления регулярных выражений был проведен анализ маркеров с целью получения их шаблонов. Например, было замечено, что всегда при наличии слова «объявляем» перед ним пишутся адресаты. Таким образом, наше регулярное выражение ищет слово «объявляем», которое может быть написано с небольшими отличиями, и проверяет несколько слов перед ним, имеют ли они окончания, например, «ымъ», «имъ», «амъ», «ему», «у» и т. д. Если все условия выполнены, то это означает, что нам удалось найти параметр «адресат» в тексте.

Другой аналогичный пример — извлечение места создания/получения. В регулярном выражении для поиска этого параметра мы используем, что они обычно следуют после слов «писана в» (далее следует город, где был написан это текст) или для второго варианта «получена в» (далее следует город, где был получен это текст), то есть мы

ищем указанные сочетания, а дальше извлекаем следующие за ними слова. Кроме того, после указания места обычно идет дата, и поэтому следующие фрагменты текста проверяются на предмет совпадения с шаблоном «<число> году <название месяца> <число> дня» или «<название месяца> <число> дня <число> году» и т. д.

Приведем несколько полученных таким образом регулярных выражений. Например, «субъект/адресант» определяется по следующему шаблону:

```
 '(?<= [о|О]т\\sвсего)\\s[А-Яа-яЁеA-Za-z]{3,}?\\s[А-Яа-яЁеA-Za-z]
{3,}?\\s'
 '(?<= [у|У]\\sce[и|й]\\sграмоты\\sнаша)\\s[А-Яа-яЁеA-Za-z]{3,}
?\\s[А-Яа-яЁеA-Za-z]{3,}?\\s'
```

То есть после фразы «от всего» следует параметр «субъект», второй шаблон ищет по фразе «у сеи грамоты наша» и затем следует наш параметр поиска. Для параметра «адресат» используется следующее выражение:

```
 '([А-Яа-яЁеA-Za-z]+(?:(?:ым[ъ|ь]?)|(?:ом[ъ|ь]?)|(?:ам[ъ|ь]?)|
(?:ему)|(?:у))[\\s]?[и|i]?\\s)?([А-Яа-яЁеA-Za-z]+(?:(?:ым[ъ|ь]?)
|(?:ом[ъ|ь]?)|(?:ам[ъ|ь]?)|(?:ему)|(?:у)))?\\s(?=[оа]?б[аяьъ]?
вляем[ъь]?)'
```

Этот шаблон ищет фразу «объявляем», перед ней проверяем слова на наличие соответствующих окончаний, если все выполнено, то параметр «адресат» найден. Параметр «жанр» в случае войсковой грамоты определяем проверкой следующих четырех шаблонов:

```
 '(?<= [с|С]ия\\sнаша)\\s[А-Яа-яЁеA-Za-z]{3,}?\\s[А-Яа-яЁеA-Za-z]
{,}\\s'
 '(?<= [с|С]ию\\sнашу)\\s[А-Яа-яЁеA-Za-z]{3,}?\\s[А-Яа-яЁеA-Za-z]
{3,}?\\s'
 '(?<= [с|С]\\sce[и|й|i]\\sнаш[а|е].)\\s[А-Яа-яЁеA-Za-z]{3,}?\\s',
 '(?<= [у|У]\\sce[и|й|i])\\s[А-Яа-яЁеA-Za-z]{3,}?\\s'
```

Первый шаблон ищет фразу «сия наша», второй — «сию нашу», третий — «с сеи нашеi», а четвертый — «у сей», и после каждой из этих фраз предполагается наличие прямого указания на жанр. Параметр «место создания» находит шаблон, ищущий фразу «писана в» и слова после нее:

```
 '(?<= [п|П]исана\\sв)\\s[А-Яа-яЁеA-Za-z]{3,}?\\s'
```

Параметр «дата создания»:

```
 ' (?<= [п|П]исана\\sв)\\s[А-Яа-яЁеA-Za-z]{3,}\\s([0-9го]{3,}?
\\sгод[уа]\\s[А-Яа-яЁеA-Za-z]{3,}?\\s)?([0-9го]{1,}?
\\s[А-Яа-яЁеA-Za-z]{3,})?'
 '(?<= [п|П]исана\\sв)\\s[А-Яа-яЁеA-Za-z]{3,}?\\s([А-Яа-яЁеA-Za-z]
{3,}?\\s[0-9го]{1,}?\\s[А-Яа-яЁеA-Za-z]{3,}\\s)?([0-9го]{3,}?
\\sгоду)?'
```

Шаблон ищет фразу «писана в», проверяет, идет ли следом параметр «место написания», и есть ли далее дата в каком-то из известных форматов написания. Аналогичные регулярные выражения составлены для всех описанных маркеров.

3. Описание приложения

В целом функционал приложения был описан выше: можно загрузить файл для анализа и получить информацию о его метаданных и жанре, если программе удалось их обнаружить по заданным шаблонам. Приведем примеры некоторых окон приложения (рис. 5, 6).

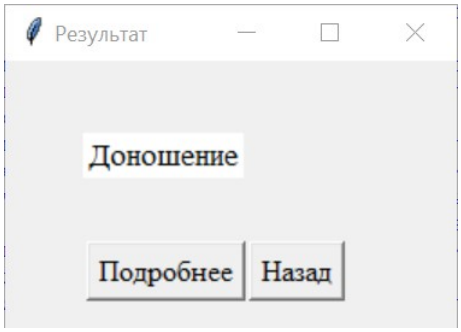

Рис. 5. Окно «Результат»

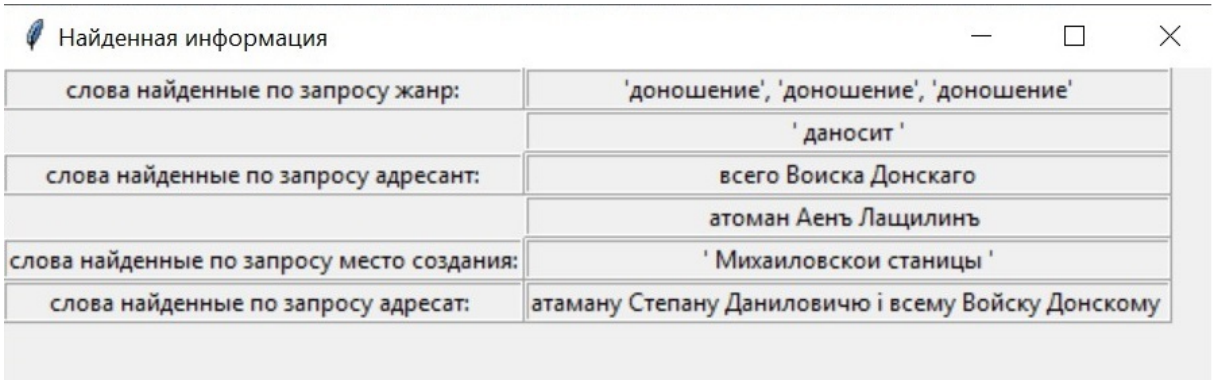

Рис. 6. Окно «Найденная информация»

Общая схема функционирования программы представлена на рисунке 7.

В настоящий момент идет апробация созданного приложения для автоматического определения видов необработанных документов, а также для их метаразметки. По результатам этого процесса алгоритмы и регулярные выражения дорабатываются и исправляются. Параллельно с этим готовятся таблицы маркеров для других типов документов, по мере их готовности программа получает новые регулярные выражения для поиска. В ближайшее время, например, будут добавлены возможности по определению метаразметки паспортов. По завершении работы над данным приложением его возможности планируется интегрировать в программное обеспечение корпуса архивных документов, работа над которым также продолжается.

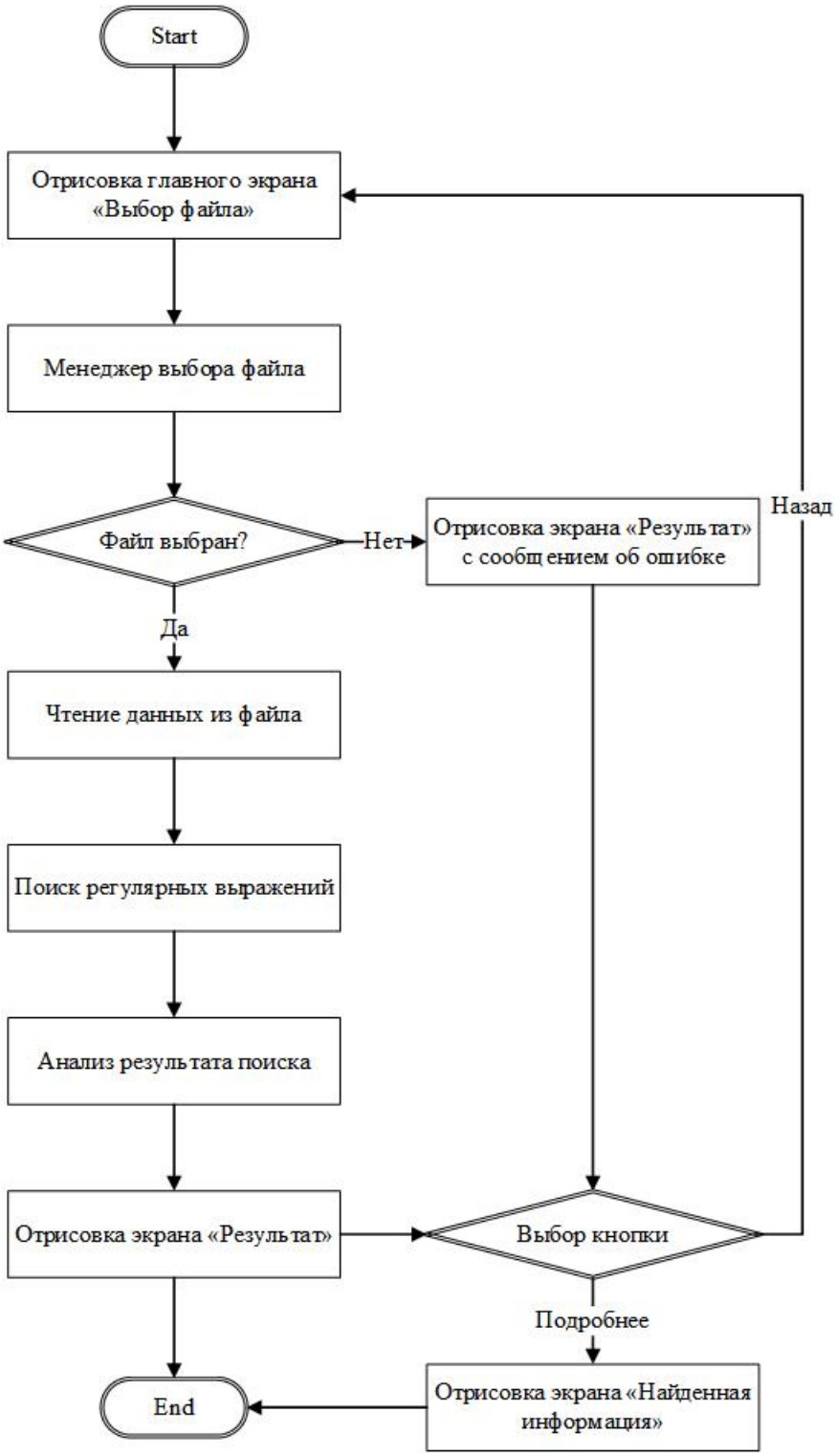

Рис. 7. Алгоритм работы приложения

ПРИМЕЧАНИЕ

[1] Работа выполнена при финансовой поддержке гранта РФФИ № 19-012-00246.

СПИСОК ЛИТЕРАТУРЫ

1. Антонова, А. Ю. Определение стилевых и жанровых характеристик коллекций текстов на основе частеречной сочетаемости / А. Ю. Антонова, Э. С. Клышинский, Е. В. Ягунова // Труды международной конференции «Корпусная лингвистика–2011». — СПб. : Изд-во С.-Петерб. гос. ун-та, 2011. — С. 80–85.

2. Барахнин, В. Б. Сравнительный анализ методов автоматической классификации поэтических текстов на основе лексических признаков / В. Б. Барахнин, О. Ю. Кожемякина, И. С. Пастушков // Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017). — М. : Федеральный исследовательский центр «Информатика и управление» Российской академии наук, 2017. — С. 252–259.

3. Горбань, О. А. Доношения и рапорты донских казаков в середине XVIII в.: источниковедческий анализ / О. А. Горбань // Вестник Волгоградского государственного университета. Серия 4. История. Регионоведение. Международные отношения. — 2019. — Т. 24, № 4. — С. 45–59. — DOI: https://doi.org/10.15688/jvolsu4.2019.4.4.

4. Гулин, В. В. Методы снижения размерности признакового описания документов в задаче классификации текстов / В. В. Гулин // Вестник МЭИ. — 2013. — № 2. — С. 115–121.

5. Епрев, А. С. Автоматическая классификация текстовых документов / А. С. Епрев // Математические структуры и моделирование. — 2010. — Вып. 21. — С. 65–81.

6. Комендантов, А. С. Автоматизация морфологической разметки архивных документов / А. С. Комендантов, А. Г. Матвеев, А. В. Светлов // Математическая физика и компьютерное моделирование. — 2019. — Т. 22, № 4. — С. 53–63. — DOI: https://doi.org/10.15688/mpcm.jvolsu.2019.4.4.

7. Косова, М. В. Параметризация текстов документов как способ жанровой идентификации / М. В. Косова // Вестник Балтийского федерального университета им. И. Канта. Сер.: Филология, педагогика, психология. — 2020. — № 1. — С. 48–55.

8. Орлов, Ю. Н. Определение жанра и автора литературного произведения статистическими методами / Ю. Н. Орлов, К. П. Осминин // Прикладная информатика. — 2010. — № 2 (26). — С. 95–108.

9. Светлов, А. В. Автоматизация процесса получения лингвистической информации: современные возможности / А. В. Светлов, А. С. Комендантов // Вестник Волгоградского государственного университета. Серия 2. Языкознание. — 2017. — Т. 16, № 2. — С. 39–46. — DOI: https://doi.org/10.15688/jvolsu2.2017.2.4.

10. Шептухина, Е. М. Жанровые параметры сказки как документа середины XVIII века в аспекте создания лингвистического корпуса / Е. М. Шептухина // Научный диалог. — 2019. — № 11. — С. 114–129. — DOI: 10.24224/2227-1295-2019-11-114-129.

11. Шептухина, Е. М. Войсковые грамоты середины XVIII века в аспекте категории модальности / Е. М. Шептухина, О. А. Горбань // Вестник Волгоградского государственного университета. Серия 2. Языкознание. — 2015. — № 5 (29). — С. 7–18. — DOI: http://dx.doi.org/10.15688/jvolsu2.2015.5.1.

12. Шептухина, Е. М. Этапы создания лингвистического корпуса войсковых грамот XVIII–XIX вв. архивного фонда «Михайловский станичный атаман» ГАВО / Е. М. Шептухина, О. А. Горбань // Гуманитарное образование и наука в техническом вузе : сб. докл. Всерос. науч.-практ. конф. с междунар. участием. — Ижевск : Изд-во Ижев. гос. техн. ун-та им. М.Т. Калашникова, 2017. — С. 428–431.

13. Cleuziou, G. On the Impact of Lexical and Linguistic Features in Genre and Domain–Based Text Categorization. / G. Cleuziou, C. Poudat // Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics. — Berlin; Heidelberg : Springer-Verlag, 2007. — P. 599–610. — DOI: https://doi.org/10.1007/978-3-540-70939-8_53.

14. Cossack Military Charters of the Mid 18th Century: Genre Distinction / O. A. Gorban, E. Yu. Ilyinova, M. V. Kosova, E. M. Sheptukhina // XLinguae Journal. — 2017. — Vol. 10, iss. 3. — P. 123–136. — DOI: 10.18355/XL.2017.10.03.10. ISSN 1337-8384.

15. Sebastiani, F. Text Categorization. / F. Sebastiani // Text Mining and Its Applications. — Southhampton, UK : WIT Press, 2005. — P. 109–129.

REFERENCES

1. Antonova A.Yu., Klyshinskiy E.S., Yagunova E.V. Opredelenie stilevykh i zhanrovykh kharakteristik kollektsiy tekstov na osnove chasterechnoy sochetaemosti [Text Collections Genre and Stylistic Categorization by Pos Co-Occurence]. *Trudy mezhdunarodnoy konferentsii «Korpusnaya lingvistika–2011»*. Saint Petersburg, Izd-vo S.-Peterb. gos. un-ta Publ., 2011, pp. 80-85.

2. Barakhnin V.B., Kozhemyakina O.Yu., Pastushkov I.S. Sravnitelnyy analiz metodov avtomaticheskoy klassifikatsii poeticheskikh tekstov na osnove leksicheskikh priznakov [Comparative Analysis of Methods of Automated Classification of Poetic Texts Based on Lexical Signs]. *Trudy XIX Mezhdunarodnoy konferentsii «Analitika i upravlenie dannymi v oblastyakh s intensivnym ispolzovaniem dannykh» (DAMDID/ RCDL'2017)*. Moscow, Federalnyy issledovatelskiy tsentr «Informatika i upravlenie» Rossiyskoy akademii nauk Publ., 2017, pp. 252-259.

3. Gorban O.A. Donosheniya i raporty donskikh kazakov v seredine XVIII v.: istochnikovedcheskiy analiz [The Donosheniya and Reports of Don Cossacks in the Mid 18th c.: Source Analysis]. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 4. Istoriya. Regionovedenie. Mezhdunarodnye otnosheniya* [Science Journal of Volgograd State University. History. Area Studies. International Relations], 2019, vol. 24, no. 4, pp. 45-59. DOI: https://doi.org/10.15688/jvolsu4.2019.4.4.

4. Gulin V.V. Metody snizheniya razmernosti priznakovogo opisaniya dokumentov v zadache klassifikatsii tekstov [Dimension Reduction Techniques for Feature Description of Documents in Text Classification Task]. *Vestnik MEI*, 2013, no. 2, pp. 115-121.

5. Eprev A.S. Avtomaticheskaya klassifikatsiya tekstovykh dokumentov [Automatic Classification of Text Documents]. *Matematicheskie struktury i modelirovanie*, 2010, iss. 21, pp. 65-81.

6. Komendantov A.S., Matveev A.G., Svetlov A.V. Avtomatizatsiya morfologicheskoy razmetki arkhivnykh dokumentov [Automation of Archival Documents Morphological Tagging]. *Matematicheskaya fizika i kompyuternoe modelirovanie*, 2019, vol. 22, no. 4, pp. 53-63. DOI: https://doi.org/10.15688/mpcm.jvolsu.2019.4.4.

7. Kosova M.V. Parametrizatsiya tekstov dokumentov kak sposob zhanrovoy identifikatsii [Parameterization of Document Texts as a Method Genre Identification]. *Vestnik Baltiyskogo federalnogo universiteta im. I. Kanta. Ser.: Filologiya, pedagogika, psikhologiya*, 2020, no. 1, pp. 48-55.

8. Orlov Yu.N., Osminin K.P. Opredelenie zhanra i avtora literaturnogo proizvedeniya statisticheskimi metodami [Determination of Genre and Author of Literary Work by Statistical Methods]. *Prikladnaya informatika*, 2010, no. 2 (26), pp. 95-108.

9. Svetlov A.V., Komendantov A.S. Avtomatizatsiya protsessa polucheniya lingvisticheskoy informatsii: sovremennye vozmozhnosti [Automation of the Process for Obtaining Linguistic Information: State-Of-The-Art Capabilities]. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznanie* [Science Journal of Volgograd State University. Linguistics], 2017, vol. 16, no. 2, pp. 39-46. DOI: https://doi.org/10.15688/jvolsu2.2017.2.4.

10. Sheptukhina E.M. Zhanrovye parametry skazki kak dokumenta serediny XVIII veka v aspekte sozdaniya lingvisticheskogo korpusa [Genre Parameters of a Narrating as a Document of the Mid-18th Century in the Aspect of Creating a Linguistic Corps]. *Nauchnyy dialog*, 2019, no. 11, pp. 114-129. DOI: 10.24224/2227-1295-2019-11-114-129.

11. Sheptukhina E.M., Gorban O.A. Voyskovye gramoty serediny XVIII veka v aspekte kategorii modalnosti [Don Cossack Army Charters of the Mid 18th Century Via the Category

of Modality]. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznanie* [Science Journal of Volgograd State University. Linguistics], 2015, no. 5 (29), pp. 7-18. DOI: http://dx.doi.org/10.15688/jvolsu2.2015.5.1.

12. Sheptukhina E.M., Gorban O.A. Etapy sozdaniya lingvisticheskogo korpusa voyskovykh gramot XVIII–XIX vv. arkhivnogo fonda «Mikhaylovskiy stanichnyy ataman» GAVO [Stages of Creating the Linguistic Corpus of Don Cossack Army Charters of the XVIII-XIX Centuries Archive Fund "Mikhailovsky Stanichny Ataman"]. *Gumanitarnoe obrazovanie i nauka v tekhnicheskom vuze: sb. dokl. Vseros. nauch.-prakt. konf. s mezhdunar. uchastiem*. Izhevsk, Izd-vo Izhev. gos. tekhn. un-ta im. M.T. Kalashnikova Publ., 2017, pp. 428-431.

13. Cleuziou G., Poudat C. On the Impact of Lexical and Linguistic Features in Genre and Domain–Based Text Categorization. *Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics*. Berlin; Heidelberg, Springer-Verlag, 2007, pp. 599-610. DOI: https://doi.org/10.1007/978-3-540-70939-8_53.

14. Gorban O.A., Ilyinova E.Yu., Kosova M.V., Sheptukhina E.M. Cossack Military Charters of the Mid 18th Century: Genre Distinction. *XLinguae Journal*, 2017, vol. 10, iss. 3, pp. 123-136. DOI: 10.18355/XL.2017.10.03.10. ISSN 1337-8384.

15. Sebastiani F. Text Categorization. *Text Mining and Its Applications*. Southhampton, UK, WIT Press, 2005, pp. 109-129.

AUTOMATION OF ARCHIVAL DOCUMENTS META TAGGING

Daniil Yu. Filimonov

Student, Institute of Mathematics and IT,
Volgograd State University
dane020597@mail.ru, matf@volsu.ru
Prosp. Universitetsky, 100, 400062 Volgograd, Russian Federation

Andrey V. Svetlov

Candidate of Physical and Mathematical Sciences, Associate Professor,
Department of Mathematical Analysis and Function Theory,
Volgograd State University
a.v.svetlov@gmail.com, andrew.svetlov@volsu.ru, matf@volsu.ru
https://orcid.org/0000-0002-8764-6132
Prosp. Universitetsky, 100, 400062 Volgograd, Russian Federation

Oksana A. Gorban

Doctor of Philological Sciences, Professor,
Department of Russian Philology and Journalism,
Volgograd State University
oa_gorban@volsu.ru
https://orcid.org/0000-0002-2345-3673
Prosp. Universitetsky, 100, 400062 Volgograd, Russian Federation

Marina V. Kosova

Doctor of Philological Sciences, Professor,
Department of Russian Philology and Journalism,
Volgograd State University
mv_kosova@volsu.ru
https://orcid.org/0000-0003-2854-8759
Prosp. Universitetsky, 100, 400062 Volgograd, Russian Federation

Elena M. Sheptukhina

Doctor of Philological Sciences, Professor,
Department of Russian Philology and Journalism,
Volgograd State University
em_sheptuhina@volsu.ru
https://orcid.org/0000-0002-8007-6042
Prosp. Universitetsky, 100, 400062 Volgograd, Russian Federation

Abstract. The main goal of this project is to create a corpus of documents from the "Mikhailovsky stanichny ataman" archival fund. The methods of corpus linguistics seem to be the most optimal in this case, since they involve the processing of a large number of texts in order to solve a wide variety of linguistic problems. Our group joined the team of philologists to provide the technical and software part of the project. The main task for us is to create a document corpus engine, that is, software that solves the tasks of storing a database of marked-up texts, executing queries to this database, and also providing users with a convenient interface for work that does not require special qualifications in the field of information technology. However, it is necessary to prepare documents for inclusion in the corpus: all texts must undergo special markup. There are many types of markup, and in the previous publications [6; 9] our group has already described the solution to the problem of morphological tagging. This article is about meta tagging. Meta tagging refers to the assignment of certain descriptive attributes to the text. In the case of office documents, these are such parameters as the type of document (genre), author (compiler), addressee, date and place of creation. Meta tagging is necessary for the implementation of the corpus search features, so that the researchers can receive text samples with specified external parameters: for example, texts of a certain type, created at a certain period, addressed to a certain addressee, etc. The archives of the "Mikhailovsky stanichny ataman" fund mainly contain documents from the Chanceries of the Don Army from the mid-18th to the first third of the 19th century, that's why there are not so many varieties of these documents. Moreover, these are mostly official documents, and they were written up according to certain templates, forms, the parameters of which can be relatively easily extracted from documents through preliminary analysis. This work is also carried out by the team of philologists from VolSU under the guidance of Professor O.A. Gorban. The result of their systematization of documents was the description of special speech markers of genre parameters for all document types in the archive. Thus, in our case, there is no need for heavy methods of statistical analysis or machine learning, it is enough to search for certain markers in the document. Moreover, the main marker in all reviewed documents is a direct indication of their type. Other markers are auxiliary elements of meta tagging. The paper is devoted to the description of the created application for determining the type of a document and its meta tagging by searching the text for certain regular expressions derived from the markers.

Key words: automation of linguistic analysis, automation of meta tagging, graphical interface, corpus-based linguistics, regular expressions.