Алгоритмы комплексного анализа русских поэтических текстов с целью автоматизации процесса создания метрических справочников и конкордансов

© В.Б.Барахнин © О.Ю.Кожемякина А.В.Забайкин
Институт вычислительных технологий СО РАН,
Новосибирский государственный университет
bar@ict.nsc.ru olgakozhemyakina@mail.ru alexey.zabaykin@gmail.com

Аннотация

В литературоведении возникает необходимость автоматизации анализа различных уровней структуры стиха, а также автоматизированного составления на основе такого анализа метрических справочников к корпусам стихов, словарей рифм и конкордансов. Целью настоящей работы является изложение алгоритмов комплексного анализа русских поэтических текстов с целью автоматизации процесса создания метрических справочников и конкордансов.

Работа выполнена при частичной поддержке РФФИ (проект 13-07-00258) и президентской программы «Ведущие научные школы РФ» (грант НШ 5006.2014.9)

1 Введение

Составление метрических справочников к корпусу стихов того или иного поэта, содержащих сведения о системах стихосложения, размерах, каталектике (ритмических окончаниях стихов), строфике, метрической композиции стихотворений, а также словарей рифм и конкордансов (алфавитных перечней всех словоформ с указанием контекстов их употребления) – важная задача литературоведения. Эти справочники и словари важны как для непосредственного изучения художественной техники поэта, так и в качестве основы для исследования влияния нижних уровней структуры стиха (метр, ритм, фонетика, лексика, грамматика) на высшие (речевой жанр, тематика, литературный жанр). Последняя задача является особенно актуальной, поскольку в этой области имеется целый ряд нерешенных проблем, некоторые из них сформулированы в [1]:

«Вопрос о том, связан ли метроритмический уровень текста с его тематикой, до сих пор является дискуссионным…

Методика выявления смысловой окраски ритма до сегодняшнего дня разработана недостаточно…

Вопрос этот [о тематических, образных и эмоциональных ассоциациях, связанных с теми или иными звуками – *авт.*] находится в стадии разработки, и пока мы не можем дать совершенно бесспорных характеристик семантики каждого звука».

Кроме того, в [1] на примере ставшей уже классической проблемы определения семантики того или иного стихотворного размера, утверждается, что методика ее решения заключается в исследовании не единичных употреблений того или иного размера, а традиций его жанрового и тематического использования, что предполагает анализ корпусов поэтических текстов. Это утверждение, очевидно, может быть отнесено и к другим проблемам исследования влияния нижних уровней структуры стиха на высшие.

Однако анализ корпусов поэтических текстов большого объема – задача чрезвычайно трудоемкая, поэтому зачастую в поле зрения исследователя попадает лишь сравнительно небольшой круг произведений поэтов-классиков, что, без сомнения, значительно снижает полноту анализируемого материала и, следовательно, достоверность полученных результатов. Таким образом, возникает необходимость автоматизации анализа различных уровней структуры стиха, а также автоматизированного составления на основе такого анализа метрических справочников к корпусам стихов, словарей рифм и конкордансов. Это позволит освободить исследователей от рутинной работы и при этом резко расширить круг изучаемых авторов.

Основные проблемы автоматизации комплексного анализа русских поэтических текстов рассмотрены нами в [2], а подходы к автоматизации процесса анализа их метрических и ритмических характеристик намечены в [3].

Целью настоящей работы является изложение алгоритмов анализа русских поэтических текстов с целью автоматизации процесса создания

Труды XVII Международной конференции DAMDID/RCDL'2015 «Аналитика и управление данными в областях с интенсивным использованием данных», Обнинск, 13-16

метрических справочников и конкордансов. Как указано в [4], такой анализ должен носить комплексный характер, чтобы, проделав однажды весьма трудоемкую работу по оцифровке корпуса текстов поэта и применяя различные программы обработки поэтических текстов, мы могли получить частотный словарь, конкорданцию, словарь рифм, каталоги метрических и строфических форм и т.п.

2 Подходы к созданию метрических словарей и конкордансов

Первые метрические справочники к стихам русских поэтов: Пушкина и Лермонтова (работа над последним не была окончена), составлены в 1930-е годы (см. обзор [4]). Естественно, эта работа велась вручную, что требовало весьма больших трудозатрат. В конце 1960-х – начале 1970-х годов, когда компьютерные технологии обработки текстов получили достаточно широкое распространение, исследования в указанной области получили новый импульс к развитию: американскими славистами были созданы словари рифм и конкордансы к стихам Пушкина, Баратынского, Батюшкова, Тютчева (для последнего – только конкорданс), советскими литературоведами – словарь рифм Лермонтова (все библиографические ссылки см. в обзоре [4]).

Разумеется, литературоведы, занимавшиеся составлением метрических словарей и конкордансов, не раскрывали детали использовавшегося ими программного обеспечения. И дело даже не в том, что для филологов этот вопрос – второстепенный. Программное обеспечение, автоматизирующее процесс составления метрических справочников, с точки зрения филолога – «ноу-хау», позволяющее получать уникальные результаты. Однако для специалистов в области компьютерной лингвистики такое программное обеспечение – непосредственный результат их научной деятельности. Так, в отделе Машинного фонда Института русского языка АН СССР был создан пакет программ UNILEX [5], предназначенный для изготовления частотных словарей, словоуказателей и конкордансов. Данный пакет был использован при создании конкорданса к стихотворениям М.Кузмина [6], при этом в статье [7] указаны его довольно существенные недостатки (отметим, что для определения количественных метрических характеристик пакет не предназначен).

Кроме конкорданса к стихотворениям М. Кузмина, за последние 25 лет были созданы Словарь языка Грибоедова [8], основную часть которого составляет алфавитно-частотный конкорданс, а также конкорданс к текстам Ломоносова [9], фактически ограничивающийся только поэтическими текстами, притом включающий лишь слова, начинающиеся на буквы А—О. Эти работы используют современные компьютерные технологии: тексты представлены на специальном языке грамматической разметки, которая основывается на «Грамматическом словаре русского языка» А.А.Зализняка, при этом предварительная грамматическая разметка корпуса выполняется при помощи программы, разработанной в компании «Яндекс», после чего проводится ручная корректировка разметки, включающая выбор вариантов разбора, снятие омонимии, разбор нераспознанных слов, исправление ошибок. В итоге размеченный корпус текстов представляет собой базу данных, с использование которой возможно исследование различных лексических, грамматических и т.п. характеристик текстов.

Отметим важную особенностей конкордансов к текстам Грибоедова и Ломоносова: в них словарные единицы сгруппированы в гнезда лексем с указанием грамматической формы каждого словоупотребления, в то время как в конкордансах к стихам Пушкина, Баратынского, Батюшкова словарные единицы суть графемы, т.е. в одно гнездо попадают и совпадающие словоформы одной лексемы, и даже омонимы и омографы, при этом, естественно, объединение словоформ по гнездам лексем не проводилось.

Итак, для конкордансов существует автоматизированная технология их создания, в которой доля ручной работы, связанной, прежде всего, с выбором вариантов разбора и устранением омонимии, довольно велика. Эта технология сравнительно легко воспроизводима, поскольку выделения графем – задача тривиальная, а грамматический разбор слов (с указанием всех возможных вариантов, выбор из которых делается вручную) можно осуществить, например, с помощью стеммера компании «Яндекс» [10].

Однако вопросы автоматизации создания метрических справочников до сих пор исследованы очень слабо. Причины этого достаточно прозрачны: если требуемые для составления конкордансов технологии обработки текстов на уровне графем, имеющие важнейшее значение для задач информационного поиска, давно разработаны и сравнительно просты, то для фонетического анализа текстов, лежащего в основе составления метрических справочников, требуются фонетические словари, включающие, как минимум, акцентуированные (т.е. содержащие ударения) и фонетически разобранные парадигмы всех слов. Так как круг задач, требующих применения таких словарей, весьма ограничен, а алгоритмы фонетического разбора и акцентуирования неоднозначны и требуют ручной корректировки результатов, то работы в этой области ведутся не слишком активно (во всяком случае, нам неизвестны словари, удовлетворяющие сформулированным требованиям). Даже наиболее полный из известных нам сетевых фонетических словарей открытого доступа – «Словарь полного фонетического разбора» [11] – содержит только начальные формы слов, поэтому необходима генерация фонетической записи словоформ. Автоматизация этого процесса не совсем

тривиальна, поскольку не существует строгих закономерностей расположения ударения в словоформах в зависимости от места его расположения в начальной форме слова.

Практически единственной работой, в которой была намечена большая программа исследований метрических, ритмических и фонетических (включая рифму) характеристик стиха, является статья [12], опирающаяся на использование системы STARLING [13]. Эта система содержит, в частности, веб-приложение для морфологического анализа [14], созданное на основе Грамматического Словаря А.А.Зализняка. Веб-приложение представляет собой морфологический анализатор, выдающий, в частности, полную акцентуированную парадигму каждого слова, имеющегося в словаре программы (к сожалению, система не позволяет генерировать парадигму произвольно заданного слова, отсутствует в ней и фонетический анализ).

Рассматриваемая программа исследований характеристик стиха была частью проекта «Автоматизированный лингвостиховедческий анализ русских поэтических текстов», которым руководил С.А.Старостин, однако после его смерти в 2005 году работы по названному проекту были свернуты.

Наконец, можно отметить сайт В.Онуфриева «Рифмовед.ру» [15], посвященный стихосложению и русской рифме, который содержит, в том числе, модуль «Экспресс-анализ стихов online», позволяющий посчитать для заданного стиха количество строф, определить их тип, установить размер стихотворения, тип рифмовки и т.п. В.Онуфриев заявляет о себе как о создателе «уникальной системы классификации русских рифм», который «открыл и объяснил новые виды русских созвучий, никем не открытые и не описанные ранее», однако точность анализа на основе его алгоритмов е слишком высока: в известном стихотворении А.Барто:

Нет, напрасно мы решили

Прокатить кота в машине:

Кот кататься не привык –

Опрокинул грузовик.

рифмовка определяется как ABCC, т.е. «решили – машине» в качестве рифмы не воспринимается, хотя это обычная неточная рифма. Особо подчеркнем: проект существует уже 13 лет, но автор не осуществил ни одной публикации в журналах, индексируемых РИНЦ, что делает практически невозможным анализ качества предложенных им алгоритмов. Теоретические же изыскания автора в области стихосложения были подвергнуты весьма резкой критике в статье [16].

3 Технология создания метрических справочников

При составлении метрических и строфических справочников целесообразно учитывать следующие двенадцать характеристик:

1. Количество строк, без учета пустых.
2. Метрика стихотворения.
3. Стопность.
4. Рифмовка строфики.
5. Количество мужских окончаний последних слов в стихотворных строках.
6. Количество женских окончаний последних слов в стихотворных строках.
7. Количество дактилических и др. окончаний последних слов в стихотворных строках.
8. Количество нерифмованных мужских окончаний.
9. Количество нерифмованных женских окончаний.
10. Количество нерифмованных дактилических и других окончаний.
11. Количество строк без конечных слов.
12. Тип строфической формы:

- стихотворения, состоящие из одной строфы (восемь строк или меньше);
- правильно повторяющиеся строфы;
- вольные стансы;
- парная рифмовка;
- вольная рифмовка.

Характеристики 1-4 учитываются в соответствии с метрическим справочником [17], характеристики 5-12 с конкордансом [18] (отметим, что их количественные значения взяты из [19]). Все перечисленные справочники созданы по стихам А.С.Пушкина, поэтому именно на них мы и тестировали излагаемые ниже алгоритмы.

Видимо, самым простым параметром для автоматического подсчета является количество строк (характеристика 1). Однако и здесь есть свои подводные камни: так, в стихотворении «Когда за городом, задумчив, я брожу...» 17-я строка по смысловым соображениям печатается в виде двух полустрок (и, естественно, именно такой вид имеет электронная версия стихотворения), но из ритмических соображений во всех справочниках эта строка считается единой, что дает расхождение при автоматическом и при ручном подсчете строк. Выявить такие особенности графического воспроизведения стихов можно при последующем анализе рифм (полустроковая структура нарушит метрику и ритм стиха), но подобная ситуация (к счастью, весьма редко встречающаяся) потребует ручного вмешательства эксперта.

Ключевой задачей при анализе поэтических текстов является определение силлабо-тонических метров (характеристики 2 и 3). Для этого

необходимо выделить стопу, состоящую из одного ударного слога в сильной позиции и одного или нескольких безударных. В зависимости от позиции ударения в стопе для двухсложных размеров различают ямб (ударение на четную позицию) либо хорей (ударение на нечетную позиции), для трехсложных размеров дактиль (ударение падает на 1-й слог), амфибрахий (на 2-й слог) и анапест (на 3-й слог).

Для автоматического определения метрической структуры поэтического текста мы воспользовались алгоритмом, описанным в [12]. Порядок работы алгоритма предполагает построение числового вектора по следующему принципу: символом 1 обозначаются безударные слоги, 2 ударные слоги односложных слов, 3 ударные слоги, занимающие первую позицию в двусложном слове, 4 ударные слоги, занимающие вторую позицию в двусложном слове, 5 ударные слоги слов, которые длиннее двух слогов. Полученный вектор анализируется по следующим правилам:

- Есть ли на нечетных позициях только символы 1 или 2? Если да это ямб.
- Есть ли на четных позициях только символы 1 или 2. Если да – это хорей.
- Есть ли на позициях номер 2, 5, 8… только символы 1, 2 или 3, на позициях номер 3, 6, 9… – только символы 1, 2 или 4? Если да – это дактиль.
- Есть ли на позициях номер 1, 4, 7… только символы 1, 2 или 4, на позициях номер 3, 6, 9… – только символы 1, 2 или 3? Если да – это амфибрахий.
- Есть ли на позициях номер 1, 4, 7… только символы 1, 2 или 3, на позициях номер 2, 5, 8… – только символы 1, 2 или 4? Если да – это анапест.
- Если 1-5 не выполнены, и отсутствует последовательность 111, это – дольник.

Характеристика 4 определяет тип рифмовки строфики. Для этого уже требуется получение фонетической информации. Фонетическая транскрипция необходима для более точного определения рифмующихся строк, нежели буквенное попарное сравнение (такие рифмы, называемые графически точными, составляют лишь небольшую долю всех рифм). Первый этап фонетической транскрипции – акцентуация – решается нами с помощью инструментария автоматической обработки текстов на естественном языке (Проект АОТ) [20], разработанного при создании системы автоматического перевода ДИАЛИНГ. Его словарь содержит порядка 3,5 миллионов акцентуированных словоформ, но, разумеется, этот словарь все равно не полон.

Для собственно фонетического анализа нами разработан модуль фонетического разбора слов, который основан на акцентуации слов с помощью последовательного (порядок важен!) применения известных правил фонетики и орфографии [21]. Следует отметить, что фонетическая транскрипция сильно зависит от ударения в слове, поэтому важно знать правильное ударение. К сожалению, это достигается не всегда из-за отмеченной выше естественной неполноты словаря ударений. Однако точность фонетического разбора в этих случаях можно повысить следующим образом. Если анализ других строф стиха (в которых проблем с акцентуацией слов не возникло) позволил нам установить его метроритмические характеристики, то на основе этих характеристик зачастую возможно установить акцентуацию слова, не входящего в словарь ударений, и провести его фонетический разбор.

Вообще говоря, задача создания более или менее полной модели русской рифмы до сих пор остается не до конца исследованной, и в настоящее время нами совместно с филологами Томского государственного университета ведется ее решение, от которого во многом будет зависеть точность определения и классификации рифм.

Для определения типа рифмовки строфики при разбиении поэтического текста на четверостишья в качестве базовых вариантов проверки выделяются кольцевая, смежная, перекрестная и сквозная рифма. В случае отсутствия названных видов строф алгоритм ищет повторяющуюся структуру длиной до 16 строк. Так, в случае поэзии Пушкина максимальная длина такой структуры – 14 строк с рифмовкой ababccddehhekk (онегинская строфа).

Характеристики 5-7, отмеченные в справочнике, – количество окончаний различных типов рифм (мужской, женской и прочих) для каждого стихотворного текста. Таким образом, не учитываются различия между дактилическими и гипердактилическими окончаниями. Для определения типа рифмы в автоматическом режиме необходимо определять ударную гласную, что осуществляется с помощью упомянутого выше словаря АОТ. Известная проблема автоматизации – невозможность выбора правильного омографа при наличии разных типов омографии (падежной, межчастеречной и др.). В случае, если в конце строки стоит слово, для которого возможны разные варианты ударений, то мы не учитываем такую строку и помечаем ее как некорректно определившуюся. Предполагается, что лингвист может в ручном режиме выбрать нужную форму омографа или, в случае отсутствия слова в словаре, произвести добавление слова в используемый тезаурус.

Характеристики 8-10 (количество нерифмованных окончаний последних слов в строке различных типов) определяются аналогично характеристикам 5-7 с учетом типа рифмовки. Если структура стих установлена, то найти количество нерифмованных окончаний не составляет особого труда. Более сложна ситуация, когда анализируемый поэтический текст относится к

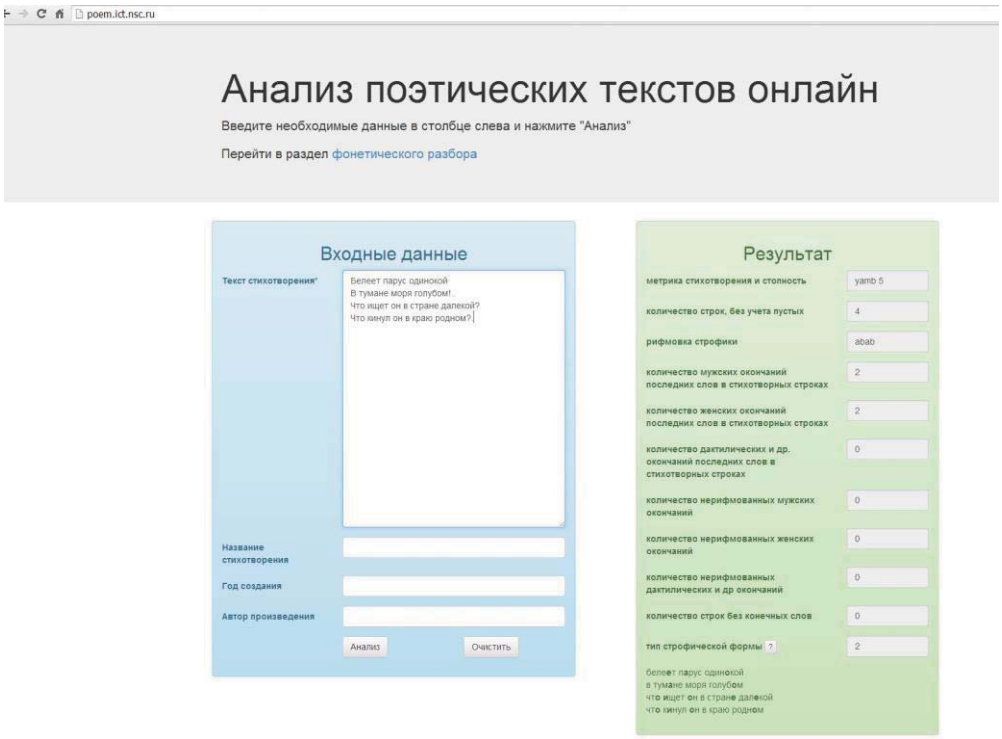

Рис. 1. Интерфейс программного средства анализа русских поэтических текстов.

разряду свободной строфики. В этом случае привязка рифмующихся окончаний ищется в некотором диапазоне, обычно не превышающем 7.

Количество строк без конечных слов (характеристика 11) определяется нахождением строк, выделяющихся из общей метрической структуры меньшим количеством слогов.

Наконец, тип строфической формы (характеристика 12) вытекает из рифмовки строфики (характеристика 4).

Что же касается программы построения конкордансов, то алгоритм, лежащий в ее основе, достаточно тривиален и аналогичен изложенному выше алгоритму из работ [8], [9]. Основная проблема – разделение омонимов (омографов) и отнесение их к нужным гнездам лексем. В настоящее время при решении этой проблемы мы не видим альтернативы работе лингвиста (на практике – достаточно грамотного носителя языка) в ручном режиме с использованием удобного программного интерфейса.

4 Практическая реализация алгоритма

Описанные выше алгоритмы реализованы на языке программирования Python 2.7 в виде программного средства обработки стихотворного текста [22]. Интерфейс программного средства представлен на рис.1. В процессе обработки стихотворения создается лог-файл, показывающий возникновение всех описанных выше случает неоднозначности, при этом в отдельную таблицу записываются слова, которые не были найдены в словаре ударений или у которых ударение неоднозначно. На основании этой таблицы лингвист может произвести добавление слова в используемый тезаурус или выбрать нужную форму омографа.

Тестирование алгоритма проводится, как сказано выше, на корпусе поэтических текстов А.С.Пушкина посредством сравнения полученных результатов с метрическим справочником [17] и с конкордансом [18]. В настоящее время точность определения характеристик составляет около 80 %, поэтому отладка алгоритма, наряду с повышением его точности (основанном, в частности, на экстраполяции ритмических характеристик строф стиха, в которых не возникло проблем с акцентуацией слов, на строфы с неакцентуированными или неоднозначно акцентуированными словами), предусматривает четкое выявление сомнительных ситуаций, для которых решение будет принимать эксперт.

5 Заключение

Изложенные алгоритмы анализа русских поэтических текстов с целью автоматизации процесса создания метрических справочников и конкордансов позволяют освободить исследователей-лингвистов от рутинной работы и при этом резко расширить круг изучаемых авторов. Планируется использование этих алгоритмов для реализации программы комплексного анализа русских поэтических текстов, представленной в нашей работе [2].

Литература

[1] Д. М. Магомедова. Филологический анализ лирического стихотворения. М.:Издательский центр «Академия», 2004.

[2] В. Б. Барахнин, О. Ю. Кожемякина. Об автоматизации комплексного анализа русского поэтического текста. Труды Четырнадцатой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2012), с. 213–217, Переславль-Залесский, 15-18 октября 2012 г.

[3] В. Б. Барахнин, О. Ю. Кожемякина, О. С. Соколова. Автоматизация процесса анализа метрических и ритмических характеристик русских поэтических текстов. Вестник Восточно-Казахстанского государственного технического университета им.Д.Серикбаева, сентябрь 2013, Вычислительные технологии, т. 18, совместный выпуск. Информационные и телекоммуникационные технологии, с.248-258.

[4] В. С. Баевский Справочные труды по поэзии Пушкина и его современников. Временник Пушкинской комиссии. АН СССР. Отделение литературы и языка. Пушкинская комиссия. СПб.: Наука, 1991, вып. 24, с. 65-79.

[5] Ж. Г. Аношкина. Лингвистический программно-источниковый пакет UNILEX+. Тексто-ориентированная компонента UNILEX-T. Бюллетень Машинного фонда русского языка, 1992, вып. 2, с. 3–7.

[6] А. В. Гик. Конкорданс к стихотворениям М. Кузмина. Т. 1-3. М.: Языки славянской культуры, 2005-2011.

[7] А. В. Бурлешин. Из песеньки слов не выкинешь... (Рецензия на книгу Конкорданс к стихотворениям М. Кузмина. Т. 1. М., 2005). Новое литературное обозрение, 2006, N 3, с. 370-384.

[8] А. Е. Поляков. Словарь языка А.С. Грибоедова. http://feb-web.ru/feb/concord/abc/

[9] А. Е. Поляков, И. А.Пильщиков, М. Б. Бергельсон. Конкорданс к текстам Ломоносова. http://feb-web.ru/feb/lomoconc/abc/

[10] Стеммер компании «Яндекс». https://tech.yandex.ru/mystem/

[11] Словарь полного фонетического разбора. http://slovonline.ru/slovar_el_fonetic/

[12] А. В. Козьмин Автоматический анализ стиха в системе Starling. Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006», с. 265-268, Бекасово, 31 мая – 4 июня 2006 г.

[13] Вавилонская Башня. Проект этимологической базы данных. Русские словари и морфология. http://starling.rinet.ru/indexru.htm

[14] Морфологический анализатор. http://starling.rinet.ru/cgi-bin/morphque.cgi?encoding=win

[15] Сайт Рифмовед.ру. http://rifmoved.ru/

[16] В. Губайловский. WWW-обозрение Владимира Губайловского. Новый мир, 2002, N 9, с. 213-216.

[17] Н. В. Лапшина, И.К. Романович, Б. И. Ярхо. Метрический справочник к стихотворениям А. С. Пушкина. М.; Л.: Academia, 1934.

[18] J. T. Shaw. Pushkin: A Concordance to the Poetry: Volumes 1 and 2. Columbus, Ohio: Slavica, 1984 / рус.пер. Дж. Т. Шоу Конкорданс к стихам А.С.Пушкина: В 2 т. М.: Языки русской культуры, 2000.

[19] J. T. Shaw. Pushkin's Rhymes: A Dictionary. Madison: Univ. of Wisconsin Press, 1974.

[20] Проект АОТ. http://nlpub.ru/AOT

[21] Правила русской орфографии и пунктуации. Полный академический справочник / Под ред. В.В.Лопатина. М: Эксмо, 2007.

[22] Анализ поэтических текстов онлайн. http://poem.ict.nsc.ru/

The Algorithms of Complex Analysis of Russian Poetic Texts for the Purpose of Automation of the Process of Creation of Metric Reference Books and Concordances

V.B.Barakhnin, O.Yu.Kozhemyakina, A.V.Zabaykin

In literary criticism there is a need to automate the analysis of different levels of the structure of the verse, as well as there is a need of computer-aided drafting on the basis of such analysis of metric guides to the poems, dictionaries of rhymes and concordances. The purpose of this paper is to present the algorithms of complex analysis of Russian poetic texts for the purpose of automation of the process of creation of metric reference books and concordances.