Программная система комплексного анализа русских поэтических текстов: модели и алгоритмы

Кожемякина Ольга Юрьевна
Барахнин Владимир Борисович

Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирск
2021

Введение

Современный подход к исследованию текстовых сообщений предполагает использование *многоуровневой модели информации*, изложенной, например, в работе немецкого исследователя В. Гитта. Нижний уровень *модели В. Гитта* соответствует шенноновскому значению термина «информация», три последующих – семиотической триаде (синтактика – семантика – прагматика), а верхний уровень носит философский характер. При этом наличие в некотором сообщении информации высокого уровня влечет за собой наличие информации всех низших уровней, но не наоборот.

Два нижних уровня сообщения (статистика и синтаксис), непосредственно связанные с кодировкой и языком сообщения, далеко не всегда влияют на верхние уровни. Так, для сообщения научного жанра практически не наблюдается зависимости понимания значения, действия и результата действия сообщения от языка, на котором написано сообщение.

Однако для некоторых типов сообщения такая зависимость весьма велика. Это относится, в частности, к сообщениям (текстам) художественного жанра, прежде всего, – к *поэтическим текстам*.

Уровни структуры стиха, подобно уровням структуры произвольного сообщения, также представляют собой *определенную иерархию*: метр, ритм, фонетика, лексика, грамматика, речевой жанр (композиционно-речевое целое), мотивика. При этом *процесс анализа стиха* предусматривает первоначальное рассмотрение каждого уровня как самостоятельной смысловой единицы с последующим связыванием этих наблюдений с другими элементами структуры.

Между уровнями структуры произвольного сообщения и стиха наблюдается *определенная корреляция*: синтаксическому уровню соответствуют метр, ритм и фонетика (согласно В. Гитту, система символов сообщения относится к именно синтаксическому уровню информации), семантическому – лексика и грамматика, создающие уровень тематики, которая, однако, частично относится и к прагматическому уровню, поскольку при анализе лирического стихотворения анализ тематики нередко включает исследование рецепции, эмоционального воздействия на читателя.

Соответствие уровней поэтического текста
уровням схемы В. Гитта

Введение

Одной из основных трудностей является необходимость *анализа корпусов поэтических текстов большого объема*. Достаточно сказать, что количество стихотворений Пушкина составляет порядка 1000. Таким образом, данная задача чрезвычайно трудоемкая, поэтому с конца 80-х — начала 90-х годов ведется работа по автоматизации анализа текста (авторы: М. Хэйвард (1996), Д. Каплан и Д. Блэй (2007), К. Боббенхаузен и Б. Хаммерих (Metricalizer2), Р. Дельмонте (SPARSAR), С.А. Старостин и А.В. Козьмин («Вавилонская башня»), wikipoetics.ru, И.А. Пильщиков).

Однако эти исследования носили разрозненный характер. Мы предлагаем *систему*, где характеристики произведения, рассматриваемые с точки зрения информационных технологий как метаданные, хранились бы вместе, с *целью удобства количественного анализа стихов*, в том числе в соответствии *с известной гипотезой К. Тарановского о зависимости высших уровней стихотворения от низших*.

Исходя из потребностей исследователей-филологов формируется одна из задач компьютерной лингвистики — *создание удобного для лингвистов программного приложения (комплекса приложений) для автоматизации работы с корпусами поэтических текстов*.

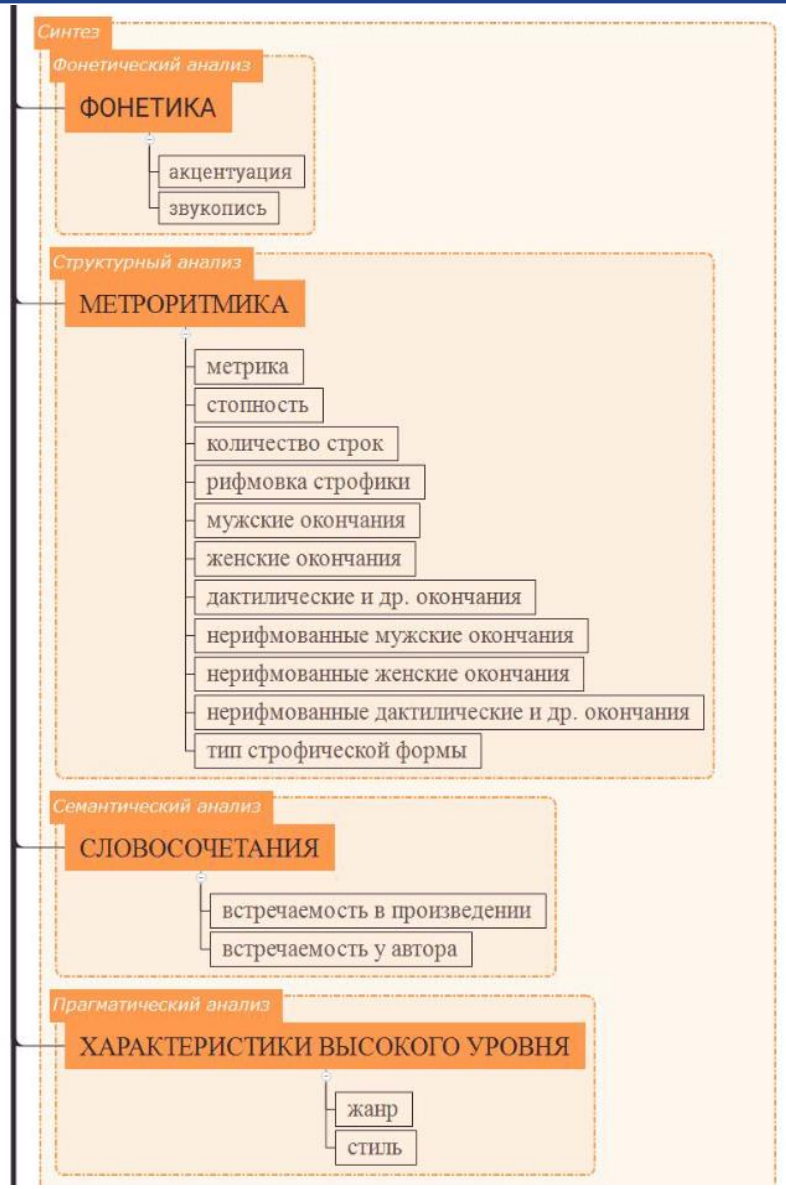
Сравнение характеристик инструментов анализа метроритмики для русского языка

Название инструмента	Анализ метра и стопности	Анализ рифмы	Наличие воспроизводимых алгоритмов, апробированных публикациями в научных журналах
«Рифмовед.ру»	+	+	–
«RitminMe»	–	+	–
«Rhymes»	–	+	–

Разработка структуры многоуровневого описания русского поэтического текста

Нами установлены категории, составляющие элементы многоуровневой структуры поэтического текста.

С точки зрения теоретической информатики, стихотворный текст содержит ряд *сущностей и категорий*. Мы выделяем следующие: акцентуация, звукопись, метрика, стопность, количество строк, рифмовка строфики, мужские окончания, женские окончания, дактилические и др.окончания, нерифмованные мужские окончания, нерифмованные женские окончания, нерифмованные дактилические и другие окончания, тип строфической формы, встречаемость в произведении, встречаемость у автора, жанр, стиль и другие.



Фонетический анализ стихотворно-текстовой информации

В процессе исследований выявилась *актуальность задачи разработки алгоритмов перевода поэтических текстов из дореформенной орфографии в современную*. При анализе фонетических характеристик поэтических текстов даже академические издания не всегда адекватно передают авторский замысел, вследствие чего требуется "контрольный" перевод из дореформенной орфографии в современную. На данный момент есть немало онлайн-сервисов по переводу из старой орфографии в современную, однако их использование в рамках полноценного программного комплекса, подразумевающего автоматическую обработку больших корпусов текстов, довольно затруднительно. Нами разработан и реализован *алгоритм перевода текстов из дореформенной орфографии в современную с учетом морфологии слов*, отличительной особенностью которого является двухэтапный перевод слова:

1) поиск и исправление правописания морфемы;
2) поиск и исправление оставшихся букв.

Второй этап подразумевает простую замену устаревшей буквы на современный ей эквивалент. Однако правописание морфем, рассматриваемых на первом этапе, не столь тривиально: далеко не всегда возможно получить верный перевод с помощью замены букв, аналогичной той, что происходит на втором этапе. Например, в слове *синія* устаревшей букве *і* соответствует современная *и*. В случае простой замены получим — *синия*, в то время как верным переводом для исходного слова является прилагательное *синие*.

Существуют два наиболее популярных алгоритма морфологического анализа для русского языка: программное приложение MyStem от компании "Яндекс" и библиотека для языка Python — *pymorphy2*. Проведенный нами анализ показал, что более точные результаты дает *pymorphy2*.

Для повышения точности работы *pymorphy2* с текстами в дореформенной орфографии нами была предложена идея, которая заключается в том, чтобы *сначала произвести временную замену окончания на его современный эквивалент*. Это приводит либо к успешному получению современной формы слова, либо к получению несуществующего или другого (нового) слова.

Рассмотрим, как реализуются эти два варианта на примере слов *тростію* и *эволюцію*:

1) на вход алгоритму перевода подается слово:

a) *тростію*,

b) *эволюцію*;

2) производится замена найденной в слове устаревшей морфемы (*-ію*):

a) *тростью*,

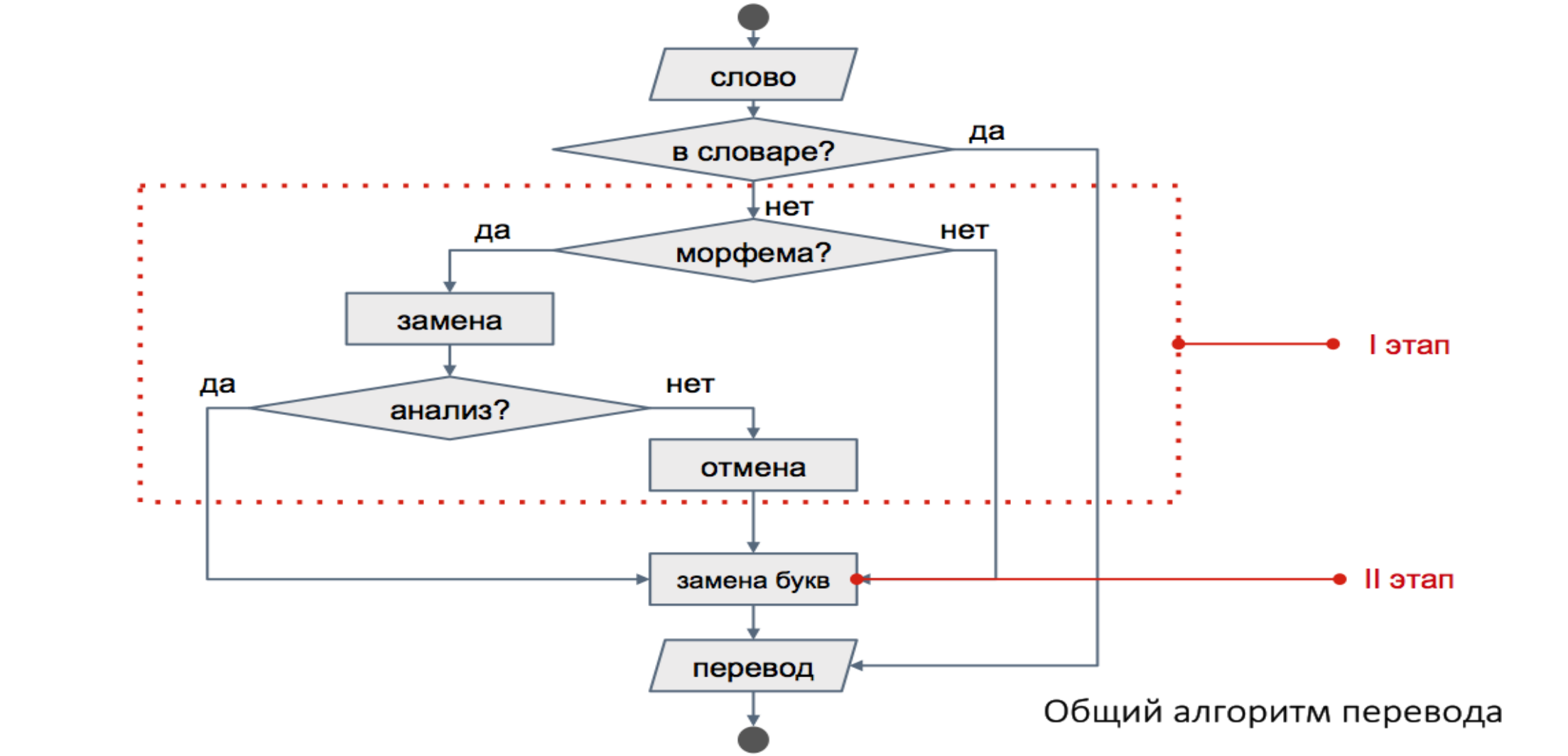
b) *эволюцью*.

При успешной замене морфологические характеристики слова соответствуют ожидаемым. То есть, слово с заменой — *тростью* — будет определено как сущ. III скл. т.п., значит перевод осуществлен. Иначе, замена приводит к изменению самого слова, а не к его переводу (*эволюцью*), поэтому характеристики будут отличаться от ожидаемых. В таком случае слово будет преобразовано к исходному виду (*эволюцію*).

Анализатор *pymorphy2* возвращает все возможные варианты морфологического разбора. Каждый из них имеет оценку вероятности того, что он является правильным. Вариант разбора, оценка которого наиболее высока, действительно оказывается верным примерно в 79% случаев. Однако этого недостаточно, поэтому *алгоритм перевода учитывает все возможные варианты*: для каждой группы слов, объединенных по морфологическому признаку, при их переводе из дореволюционной орфографии в современную мы определяем точную последовательность шагов алгоритма, учитывающую особенность изменения правописания приставок и разнообразных падежных окончаний.

В ходе тестирования была рассчитана точность перевода в режимах работы с различными текстами, которая оказалась близка к *100 %* (исключение составил перевод *концемъ* -> *концем*, ошибка в котором связана с неформализованным правилом: раньше звук [ц] был мягким).

Алгоритма перевода в современную орфографию



Общий алгоритм перевода

Исследования метро-ритмических характеристик текста

Наиболее сложной проблемой, возникающей при анализе нижних уровней поэтического текста, является довольно часто возникающая *невозможность однозначно поставить ударение с помощью базы данных акцентуации словоформ русского языка*, созданной на основе словаря А.А. Зализняка (омографы, собственные имена, «авторские» неологизмы и т.д.), а также *пропуск схемных ударений (пиррихии), наложение сверхсхемных ударений (спондеи), переносы ударений на предлоги (проклитики) и др.*

Метроритмические характеристики поэтических текстов определяются нами на основании *алгоритма из статьи [Бойков В.Н., Каряева М.С., Соколов В.А., Пильщиков И.А. Об автоматической спецификации стиха в информационно-аналитической системе* // CEUR Workshop Proceedings. 2015. V. 1536. P. 144-151]. Суть данного алгоритма заключается в *сопоставлении ритмических вариантов стиха изучаемого поэтического текста с набором ритмических шаблонов из определенного репертуара метроритмических вариантов стиха*. Основной сложностью при практической реализации этого алгоритма является то, что он *предполагает «идеальную» акцентуацию слов и совершенно не учитывает реальные проблемы автоматизированной* акцентуации, изложенные выше.

Для разрешения этих ситуаций был предложен *алгоритм расстановки ударений методом «по аналогии»*. *Суть метода* заключается в следующем: строки и строфы с неоднозначной расстановкой ударения сравниваются со строками и строфами, в словах которых ударения расставляются однозначно, и производится выбор ударения, обеспечивающего единство метрической характеристики для всего стихотворения. Для реализации этого метода при переводе поэтического текста в последовательность символов «*c*» и «*C*» (обозначающих безударный и ударный слоги соответственно) в словах с неоднозначной расстановкой ударения позиции всех возможных вариантов ударения обозначаются символом «*x*». Таким образом, *текст представляется в виде таблицы символов «c», «C», «x» размерности n (число строк поэтического текста) на m (строка с максимальным количеством слогов).*

Далее, *для устранения неоднозначности расстановки ударения* в каждой строке таблицы осуществляется поиск элемента «*x*» и взятие столбца таблицы по индексу этого элемента. В этом столбце ищется максимально часто встречающийся однозначный элемент и его значение присваивается элементу «*x*». Для случаев нарушения метрического ударения, связанного с проклитикой, была составлена БД проклитик на основе отдельных указаний в словарных статьях словаря А.А. Зализняка. В ней содержится информация о вариантах акцентуации сочетаний некоторых слов и предлогов. *Текст анализируется на наличие предлогов*. Если в тексте встречается предлог, то осуществляется поиск сочетания этого предлога в объединении со словом, стоящим от него справа. При обнаружении данного сочетания в БД проклитик, из нее извлекается *информация о вариантах акцентуации данного сочетания*. В случае неоднозначных вариантах расстановки ударений мы вновь прибегаем к методу «по аналогии».

Исследования метро-ритмических характеристик текста

Предложены *две модифицирующие поправки к алгоритму из статьи* [Бойков и др., 2015], которые позволили *снизить чувствительность алгоритма к входным параметрам, а также учесть пиррихии и спондеи* (соответствующие математические формулы приведены в работе [Barakhnin V.B., Kozhemyakina O.Yu., Kuznetsova I.V. Development and Implementation of the Algorithm for Automatic Analysis of Metrorhythmic Characteristics of Russian Poetic Texts CEUR Workshop Proceedings (2019 г.)].

Модифицированный алгоритм определяет метр и стопность на корпусе стихотворений Пушкина с точностью в 95,5 %, при этом основной процент погрешности дают стихотворения с отточиями: неоконченные, с обсценной лексикой и т.п.

Автоматическое определение характеристик, связанных со стопностью, требует нахождения рифмующихся строк. *Алгоритм поиска рифм основан на соображениях возможности их образования:* строки рифмуются, если у последних слов в строке одинаковая позиция ударного слога и фонетически совпадают окончания. Для выявления фонетически совпадающих окончаний использовались данные об окончаниях из работы [Жирмунский В.М. Рифма, ее история и теория. Пг.: Academia, 1923]. В ней приведены пары буквосочетаний, отражающих звуки рифмующихся стиховых окончаний из литературы XVIII-XIX веков (разумеется, даже для литературы указанного периода упомянутый список окончаний является далеко не полным).

В ходе тестирования этой версии алгоритма было выявлено, что *часть ошибок в определении типа строфики связана с неполнотой множества фонетически рифмующихся окончаний*, поэтому алгоритм претерпел некоторые *модификации для повышения точности*.

На основании перечисленных алгоритмов была разработана и реализована в виде веб-приложения *программа по определению метроритмических характеристик*, которая выдает следующие *данные о тексте*:

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.

Выходные данные веб-приложения могут заполнять таблицы, аналогичные основным таблицам I, II, VII, X, XIII, XIV, XV в классическом метрическом справочнике [Лапшина Н. В., Романович И. К., Ярхо Б. И. Метрический справочник к стихотворениям А. С. Пушкина. М.; Л.: Academia, 1934], а также приведенные в нем индексы: алфавитный, строф и размеров.

Фактуры

Базисной единицей анализа характеристик фоно-метрического уровня поэтического текста являются так называемые *фактуры*, учитывающие строфику и метроритмику. Согласно определению из "Поэтического словаря" А.П. Квятковского, фактура стиха — это "совокупность технологических приемов поэта, которые наряду с идейно-художественным замыслом определяют степень его поэтической культуры и мастерства", которая "проявляется главным образом в лексике, фонике, ритме стиха". Мы используем "минимальное" определение фактуры, в котором учитываем только метрические и строфические характеристики. Однако более тонкий анализ поэтических текстов предполагает, что в рассмотрение должны приниматься также ритмические характеристики, учитывающие, в дополнение к "идеальной" метрике, еще и пиррихии, спондеи и т. п., а также тонкие фонетические характеристики в рифме: точность, открытость, исследуемые фоносемантикой звуковые характеристики и т. д.

Итак, *фактура* - *совокупность метроритмических и строфических характеристик поэтического текста*, таких как:

1. Стихотворный размер:
   - стопность,
   - вид размера (метр).
2. Схема строфы (строфика):
   - перечисление групп рифмующихся строк с учетом типа клаузул.

Метроритмические характеристики учитывают размер строфы, стопность каждой строки, при этом полнота/неполнота строк по умолчанию входит в строфические характеристики. Под *строфическими характеристиками* мы понимаем вид взаимного расположения рифмических цепей (смежный перекрестный, охватный т.д.) и количество слогов, объединенных ударением (что определяет рифму мужскую, женскую, дактилическую и т.д.) *Фактура стихотворения в целом* — это *фактура входящих в него строф*, в предположении, что оно равнострофическое. Фактура является структурной моделью поэтического текста и однозначно определяет метроритмический и строфический шаблон стихотворного текста.
Нами построена *математическая модель метра и стопности стихотворения* на основе модифицированной нами методики И. А. Пильщикова, посредством учета возможности неоднозначной расстановки ударений, связанной с наличием омографов, проклитик, пиррихиев и т. п. На основании дополненной модели разработан и программно реализован *модифицированный алгоритм определения метра и стопности*.
Построена и программно реализована *математическая модель автоматического определения строфики стихотворения*.
На основании данных "Метрического справочника" (Лапшина Н.В., Романович И.К., Ярхо Б.И. Метрический справочник к стихотворениям А.С. Пушкина. М.; Л.: Academia, 1934), *составлена полная таблица примеров фактур, используемых А.С. Пушкиным в равностофических монометрических поэтических текстах*. Разработан *алгоритм определения фактуры*, на его основе реализована программа на языке Python, определяющая фактуру анализируемого стихотворения.

Составление конкордансов и справочников, извлечение словосочетаний, составление словарей языка поэтов

Для *работы по созданию, корректировке и пополнению метрических справочников* предполагается *две роли пользователей*, возможности которых определяются прохождением пользователем авторизации в веб-приложении. Неавторизованный пользователь может составлять справочник, но не может вносить в него изменения: предусмотрена только демонстрация базовой функциональности веб-приложения. Авторизованный в системе пользователь имеет возможность как загружать корпус текстов, так и работать с ним: вносить изменения и сохранять их в системе для дальнейшего использования. Если система не может корректно определить метроритмические характеристики поэтического текста, то она относит его к дисметрическим, после чего такие стихотворения поступают на анализ к эксперту.

Современный подход к построению конкордансов (например, Конкорданс к стихам А.С. Пушкина, Ф.И. Тютчева, М.А. Кузмина) состоит в том, что *словарная статья представлена основным заголовочным словом* - лексема, приведенная к лемме (т.е. снята неоднозначность в случае омонимов, омографов и полисемии), внутри которой также разделяются все возможные омоформы. Для разделения омоформ между собой они могут быть сопровождены грамматической информацией. Каждая словарная статья (ее заголовочное слово) может быть дополнена кратким толкованием для учета омонимии и полисемии.

Основной трудностью автоматизации построения конкордансов является *разрешение неоднозначных графем*, то есть определение их лексических и грамматических характеристик. Для решения этой задачи нами составлена *классификация различных случаев неоднозначности, обусловленная комбинацией проблем филологии и компьютерной лингвистики* и необходимая при реализации системы автоматизированного анализа текстов, в том числе при составлении словарей поэтического языка и конкордансов, основанных на совпадении и несовпадении ударений и, в свою очередь, постоянных и непостоянных грамматических признаков.

Предложен *алгоритм автоматического построения конкордансов с фиксацией неоднозначности и снятием неоднозначности акцентуирования*. Затем зафиксированная неоднозначность может быть устранена экспертом-пользователем.

Исследование жанровых и стилевых характеристик поэтических текстов

Нами предложены *принципы формирования обучающих выборок для алгоритмов определения стилей и жанровых типов*. Обоснован *совместный («двумерный») классификатор жанровых типов и стилистической окраски поэтических текстов, основанный на определении взаимозависимости жанрового типа и стилистической окраски*. Проанализированы *принципы формирования обучающих выборок для алгоритмов определения стилей и жанровых типов*, в том числе с использованием наиболее известных приемов ансамблирования базовых алгоритмов в композиции, таких, как взвешенное голосование, бустинг и стекинг, причем в качестве характеристических признаков стихотворений использовались одиночные слова, биграммы и триграммы. Предложенные алгоритмы существенно облегчают работу эксперта при определении их стилей и жанров путем предоставления соответствующих рекомендаций.

Эксперимент с определением стиля

Метод	SVM, нейтральный игнорируется	SVM	SVM, rbf ядро	Многослойная нейросеть	Логистическая регрессия	Линейная регрессия, нейтральный игнорируется	Линейная регрессия
Среднее	0,76	0,80	0,62	0,77	0,76	0,70	0,70
Max	0,92	0,96	0,85	0,96	0,96	0,82	0,80
Min	0,58	0,57	0,12	0,46	0,46	0,45	0,58
Низкий	0,723	0,70	0,31	0,33	0,72	–	–
Средний	0,86	0,86	0,75	0,96	0,85	–	–
Высокий	0,0	0,0	0,10	0,0	0,0	–	–

Эксперименты по определению жанра и стиля

Триграммы + SMOTE для определения стиля

Классификатор	Среднее	Max	Min
SVM AdaBoost	0.98	0.99	0.98
XGBoost	0.93	0.94	0.92
Многослойный персептрон	0.99	0.99	0.99
Голосование, hard	0.98	0.99	0.98
Голосование, soft	0.98	0.99	0.98
Стекинг	0.98	0.99	0.99

Триграммы + случайное дублирование миноритарных классов для определения жанра

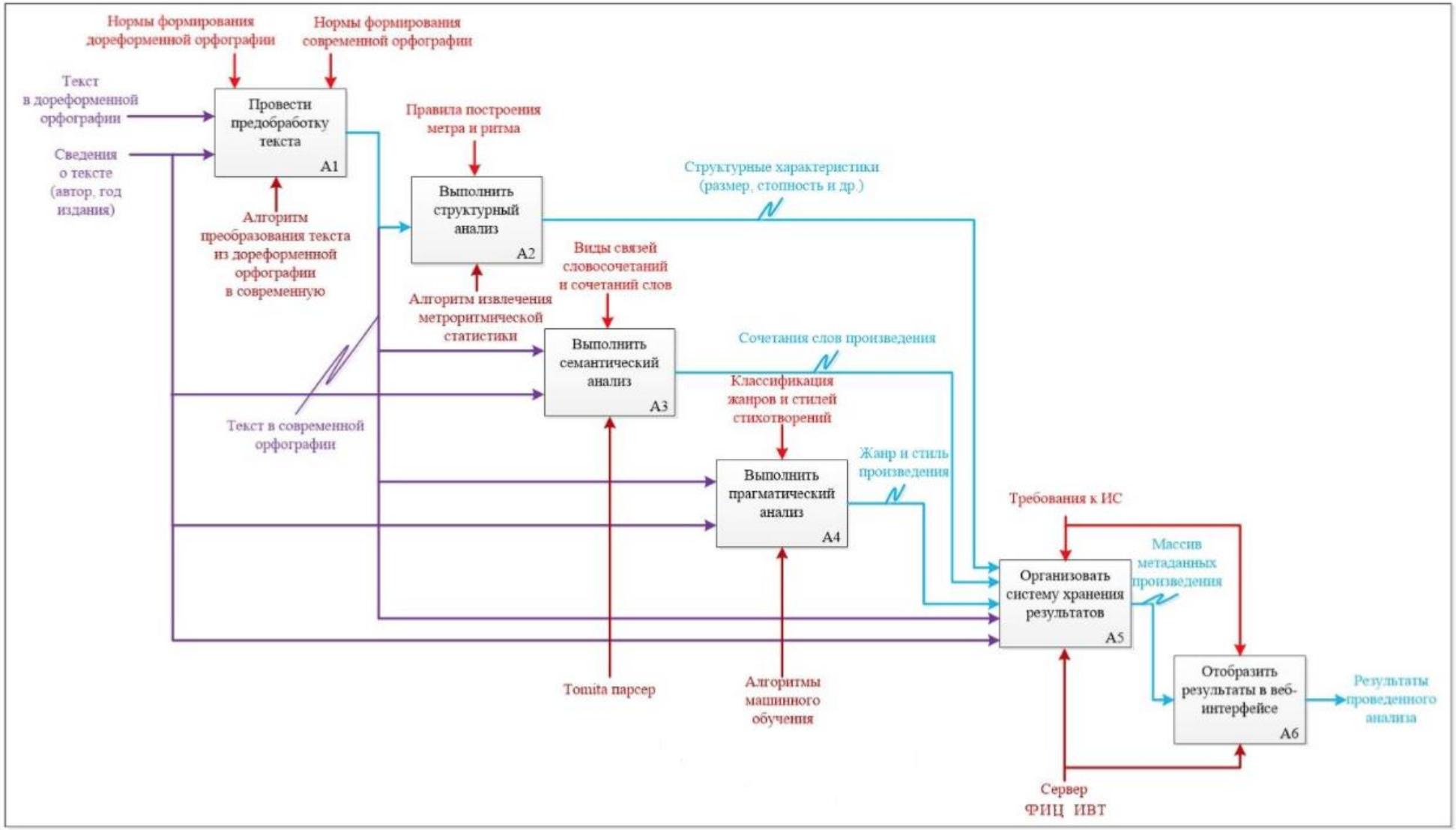
Классификатор	Среднее	Max	Min
SVM AdaBoost	0.95	1.00	0.86
XGBoost	0.94	1.00	0.84
Многослойный персептрон	0.97	0.99	0.95
Голосование, hard	0.96	1.00	0.91
Голосование, soft	0.96	1.00	0.91
Стекинг	0.96	1.00	0.88

Несмотря на широкое распространение, модели машинного обучения остаются в основном *«черными ящиками»*. Пользователям недоступна информация о причинах предоставленных прогнозов, хотя такие знания имеют *большое значение при дальнейшем анализе результатов*. Основная проблема заключается в *недоверии пользователей модели или предсказанию*, что приводит к отказу от ее неиспользования. Эти проблемы возможно преодолеть путем создания программной *системы*, позволяющей загружать *категоризированные данные*, а на выходе отображать *подробный отчет о результатах классификации*.

Нами предложены *алгоритмы программной системы автоматического определения стилей и жанровых типов поэтических текстов*, которая позволяет экспертам в области филологии удобным способом загружать необходимые данные, выбирать автоматический классификатор и полноценно анализировать результат классификации, основываясь на его обосновании, полученном с помощью алгоритма LIME и его реализации в библиотеке ELI5. Разработанная программная система *существенно облегчает работу эксперта* при определении стилей и жанров поэтических текстов *благодаря рекомендованным системой результатам*. Помимо этого, достигается минимизация времени обучения взаимодействия пользователя с системой, что приводит к увеличению скорости и качеству классификации. Следует отметить, что данное решение не привязано к каким-то конкретным категориям классификации, что делает её универсальной.

Низкий стиль Признаки с наибольшим весом		**Средний стиль** Признаки с наибольшим весом		**Высокий стиль** Признаки с наибольшим весом	
Вес	**Признак**	**Вес**	**Признак**	**Вес**	**Признак**
+ 1.991	ты	+ 1.702	он	+ 1.522	из словаря русского языка XVIII в.
+ 1.925	которого	+ 0.702	друг	+ 1.518	роковой
10174 признака с положительным весом		*11710 признаков с положительным весом*		*15007 признаков с положительным весом*	
25605 признаков с отрицательным весом		*24069 признаков с отрицательным весом*		*20772 признака с отрицательным весом*	
– 1.686	вас	– 1.379	<BIAS>	– 1.764	которого
– 10.453	<BIAS>			– 2.171	чего
				– 5.013	<BIAS>

Алгоритм обработки поэтического текста



Концептуальное описание системы

Система описывает следующие *типы сущностей*:

- *Авторы*
- *Тексты*
- *Лексемы*
- *Рифмы*
- *Метроритмические характеристики (фактуры)*
- *Жанры*
- *Стили*

Центральным компонентом является *система хранения текстов и их метаданных*. *Базовый элемент описания* - слово в каждом своем вхождении в поэтический текст. Предлагается *четырехуровневое кодирование*:

1. ID Автор.
2. ID Стихотворение.
3. ID Строка в стихотворении.
4. ID Слово в строке.

Таким образом, каждое слово определяется четырьмя индексами.

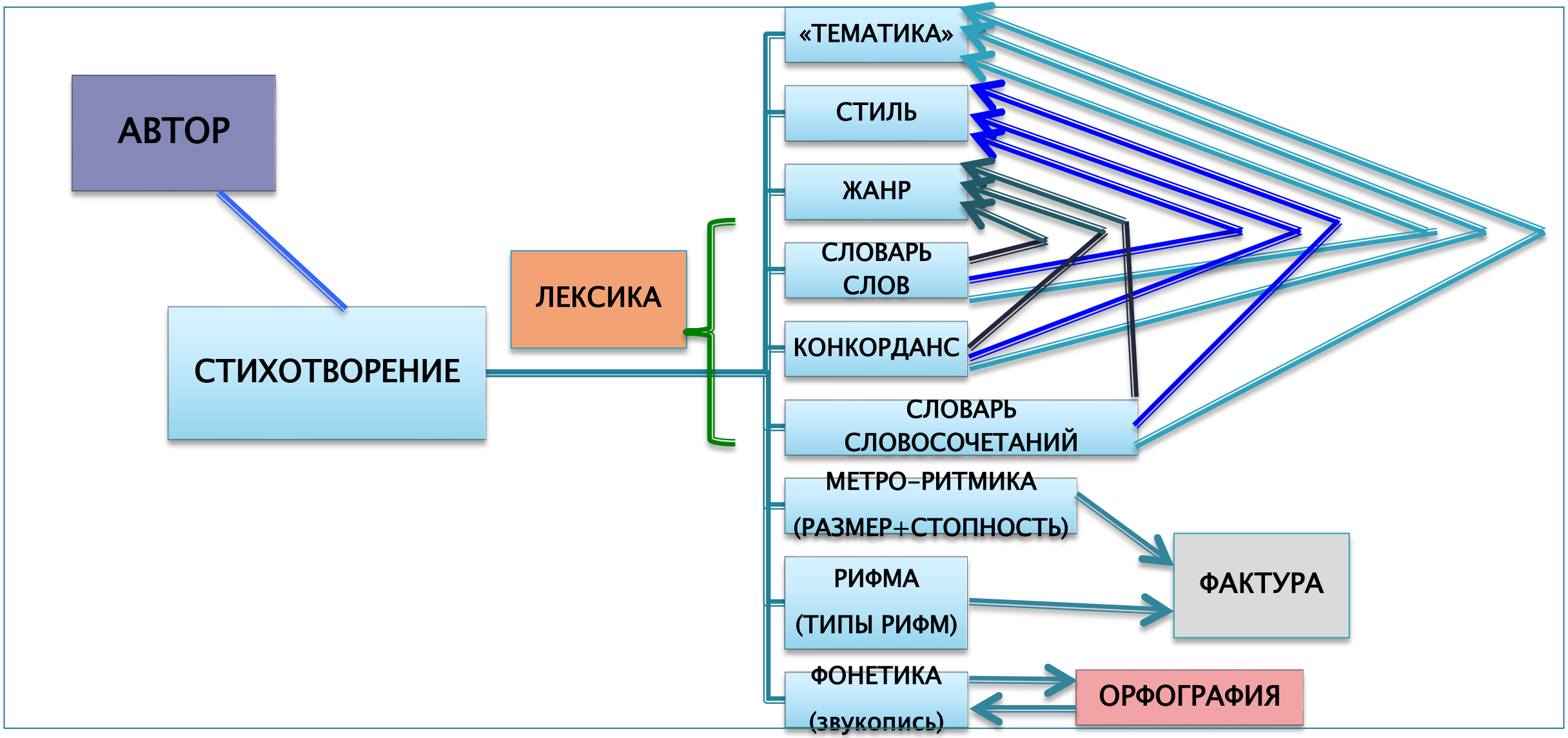
Система позволяет *автоматически генерировать следующие словари и справочники*, основные для филолога-стиховеда.

- *Лексический словарь*
- *Словарь языка*
- *Конкорданс*
- *Словарь рифм*
- *Метроритмический справочник*

В лексическом словаре хранятся *слова в начальной форме*. Обращение к начальным формам слов позволяет получить *словарь языка*. Обращение к словоформам в контексте слов – *конкорданс*.

Информация о каждой рифме представлена в виде n-арного отношения, где $n \geq 2$. Посредством обращения к этой компоненте организуется *автоматическое составление словаря рифм*.

Обращение к метроритмическим и строфическим метаданным корпуса стихотворений поэта позволяет организовать *метроритмический справочник*.


АВТОР
СТИХОТВОРЕНИЕ
ЛЕКСИКА
«ТЕМАТИКА»
СТИЛЬ
ЖАНР
СЛОВАРЬ СЛОВ
КОНКОРДАНС
СЛОВАРЬ СЛОВОСОЧЕТАНИЙ
МЕТРО–РИТМИКА (РАЗМЕР+СТОПНОСТЬ)
РИФМА (ТИПЫ РИФМ)
ФОНЕТИКА (звукопись)
ФАКТУРА
ОРФОГРАФИЯ

Таким образом:

1. Обосновано соответствие между уровнями многоуровневой модели информации и уровнями структуры поэтического текста.
1. Разработана структура многоуровневого описания русского поэтического текста.
1. Разработаны алгоритмы фонетического анализа стихотворно-текстовой информации.
1. Исследованы принципы и разработаны алгоритмы анализа метро-ритмических и строфических характеристик текста (фактур).
1. Исследованы принципы и разработаны алгоритмы составления конкордансов и справочников, извлечения словосочетаний, составления словарей языка поэтов.
1. Разработаны алгоритмы автоматизированного извлечения жанровых и стилевых характеристик.
1. Разработана модель программной системы комплексного анализа поэтических текстов.

СПАСИБО ЗА ВНИМАНИЕ!