Об автоматической спецификации стиха в информационно-аналитической системе

В. Н. Бойков, М. С. Каряева, В. А. Соколов В. А., И. А. Пильщиков

Ярославский государственный университет им. П. Г. Демидова
Ярославль
Московский государственный университет им. М. В. Ломоносова
Москва
boykov_bh@bk.ru; mari.s.ka@mail.ru; valery-sokolov@yandex.ru; pilshch@yandex.ru

Аннотация

В работе рассматриваются такие автоматические процедуры спецификации поэтического текста, как метрико-ритмическая разметка и идентификация стихового метра. Алгоритмизация процедур идентификации стиховых размеров такого метра как «ямб» показывает перспективность выработанного подхода для всего спектра систем русского стихосложения.
Работа поддержана Российским фондом фундаментальных исследований, грант №13-06-00448.

1 Введение

Информационно-аналитическая система русской поэзии (ИАСРП), разрабатываемая как открытый сетевой ресурс http://wikipoetics.ru/, представлена двумя основными компонентами: проблемно-ориентированным «Тезаурусом по поэтологии» (около 2000 терминов) и «Блоком анализа и спецификации» текстовых объектов [1,3-5]. Структура ИАСРП представлена на Рис. 1.

В «Блоке анализа и спецификации» указанной системы выделяются два основных программно-алгоритмических комплекса задач: спецификация терминологических статей тезауруса и спецификация поэтического произведения (ПП).

Структура комплекса автоматических процедур метрико-ритмической спецификации ПП наглядно представлена на Рис. 2., где выделены следующие группы автоматических решений:

- метрико-ритмическая разметка текста;
- заполнение полей спецификации ПП;
- идентификация метра.

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

Одной из существующих систем, позволяющих определять спецификации стиха, является система SPARSAR (System for Poetry Automatic Rhythm and Style AnalyzeR) [17-18]. SPARSAR производит анализ стихотворений на разных уровнях: на уровне предложения, строки и строфы. Для поиска ритма последовательно проводится синтаксический, семантический и грамматический анализ. Алгоритм оценивает как рифмы на уровне строфы, а затем и на всем стихе. Для анализа используется тезаурус WordNet [19], из которого можно извлечь дополнительную информацию, по словам из исследуемого стиха. В качестве методов по определению ритма была взята работа Малкольма Хейворда [16], позволяющая анализировать английский поэтический метр с использованием подхода под названием коннекционизм.

2 Поля спецификации ПП

На основе опыта поэтологических исследований и имеющихся традиций лингвостиховедческой разметки стиха была выделена следующая совокупность полей спецификации ПП:

- автор1;
- пол автора1 (М, Ж);
- год рождения автора;
- автор2, 3 и т.д. (в случае наличия соавторов);
- пол автора2, 3 и т.д.;
- год рождения автора2, 3 и т.д.;
- год создания стихотворения (диапазон либо точная дата);
- заглавие1;
- incipit (первая строка);
- заглавие2, 3 и т.д. (альтернативные заглавия, если есть);
- incipit заглавия2, 3 и т.д.;
- атрибутивный статус (несомненное vs. приписываемое);
- цикл (принадлежность к циклу, вхождение в состав цикла);
- книга стихов (вхождение в состав книги стихов);
- метр1 (вид акцентного стиха);

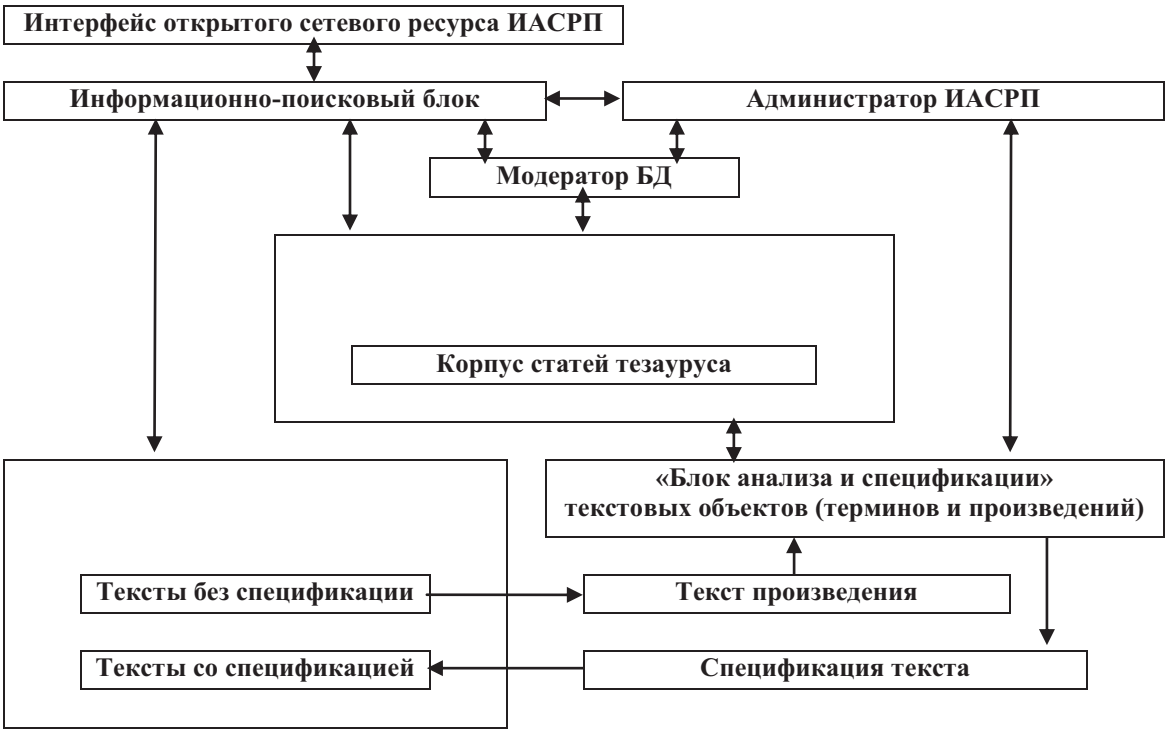

Рис. 1. Структура открытого сетевого ресурса ИАСРП

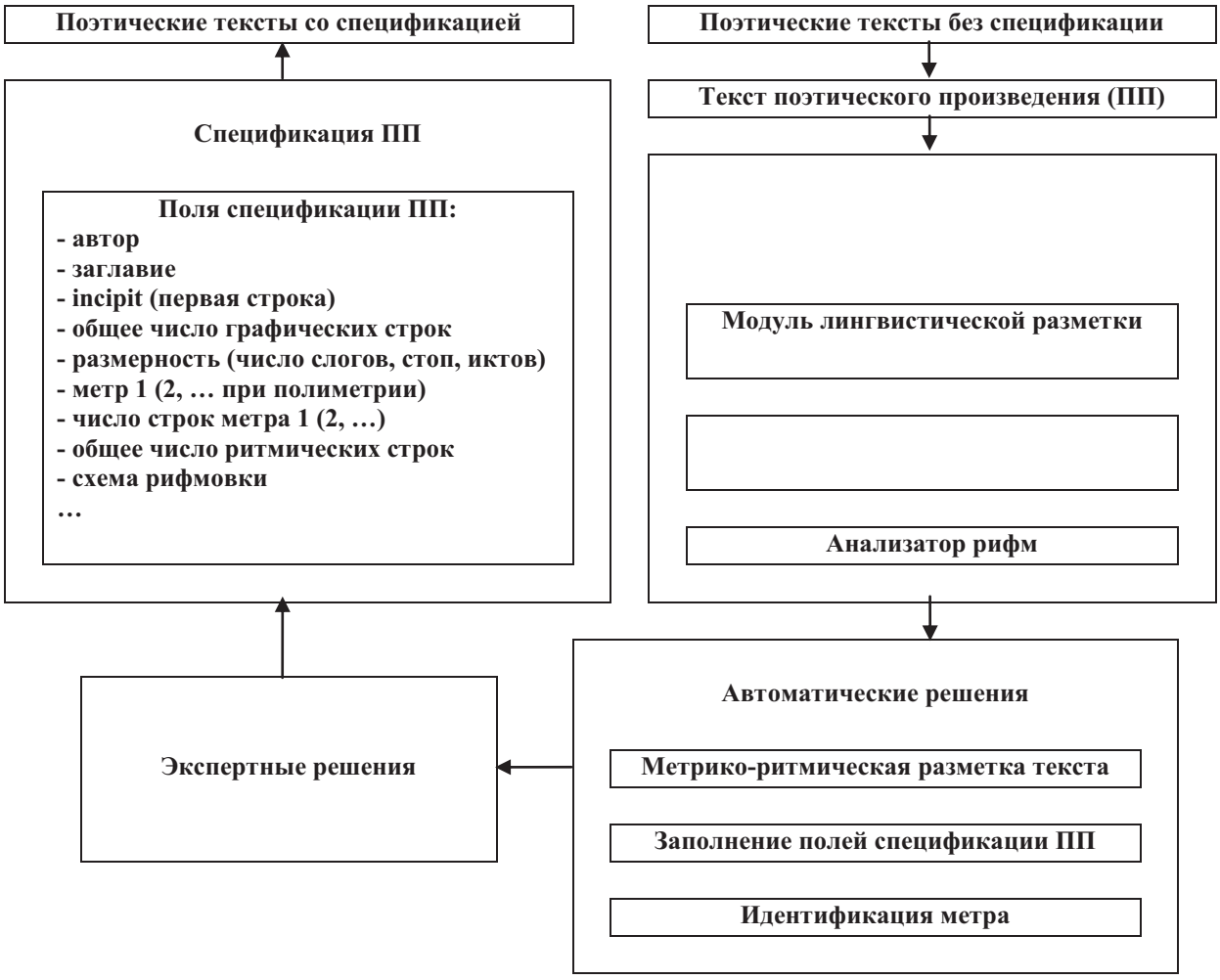

Рис. 2. Структура комплекса автоматических процедур спецификации ПП

- размерность1 (число слогов, стопность, иктовость, число междуиктовых словоразделов);
- число слогов анакрузы (0, 1, 2, 3,…);
- число безударных слогов между иктами);
- число безударных слогов клаузулы (1, 2, 3. 4,…);
- метр2,3 и т.д. (для полиметрии и смешанных размеров);
- число строк метра1;
- число строк метра2,3 и т.д.;
- общее число ритмических строк;
- общее число графических строк (в обычном случае равно числу ритмических строк);
- схема рифмовки;
- строфика (включая твердые формы);
- длина строфы (число строк);
- число строф;
- язык стихотворения (по умолчанию: русский);
- язык(и) фрагментов стихотворения (при наличии в тексте фрагментов на других языках);
- язык оригинала (для переводных и пародийных произведений);
- заглавие оригинала (для переводных и пародийных произведений);
- автор(ы) оригинала (для переводных и пародийных произведений);
- пол автора оригинала;
- год рождения автора оригинала;
- дополнительные характеристики стихотворения (здесь можно при необходимости указывать существенные, но «нестандартные» характеристики: акростих, фигурное стихотворение, особая графика и т.д.).

Для названного комплекса автоматических процедур метрико-ритмической спецификации ПП из приведенной выше совокупности полей спецификации ПП выделяются две группы полей метрико-ритмической спецификации:

Размерность строки (стиха) P_c:

- число слогов в строке (сложность);
- число ударных слогов – иктов (иктовость);
- число междуиктовых словоразделов;
- число слогов анакрузы (0, 1, 2, 3,…);
- число слогов клаузулы (1, 2, 3, 4,…);
- размерность междуиктовых интервалов (число безударных слогов между иктами);
- метр строки (стопа);
- размерность метра (стопность);
- тип акцентного стиха и размерность (иктовость).

Размерность произведения:

- общее число графических строк (в обычном случае равно числу ритмических строк);
- общее число ритмических строк;
- число рифм;
- номера строк с одинаковой рифмой;
- схема рифмовки;
- строфика (включая твердые формы);
- длина строфы (число строк);
- число строф.

Последние три поля в данном случае не используются.

3 Метрико-ритмическая разметка текста и заполнение полей спецификации ПП

Важнейшей частью комплекса спецификации ПП является подготовка его текста с помощью модуля лингвистической разметки к метрико-ритмическому анализу, для корректного проведения которого необходимо дать конструктивное определение термина ПП – «стихотворение».

Под стихотворением понимается разбитый на строки независимо от синтаксических связей в тексте и представленный набором таких строк орфографический текст. Каждая строка (Ст) состоит из последовательности символов – букв русского алфавита и дефиса, набора знаков пунктуации, словоразделов и конечного символа (специального символа, обозначающего конец строки).

Комплекс задач автоматизированной лингвостиховедческой разметки достаточно подробно рассмотрен в работах [10, 12], где заострено внимание на проблемах акцентологической разметки. Расстановка ударений зависит от снятия некоторых видов омонимии, и при не снятой омонимии возникает вариативность акцентуации стихотворной строки. Кроме того, усложняет такую разметку способность вполне ударяемых слов атонироваться и безударных слов принимать ударение. В данном случае акцентологическая разметка опирается на «Грамматический словарь» Зализняка [8] (в электронный формат переведен Старостиным [11,14]), где приводятся правила акцентуации для каждой парадигмы.

Автоматизация метрико-ритмической разметки стихотворения включает в себя следующие процедуры.

1. Нумерация строк ПП с предварительной их разметкой с помощью конечного символа ПП=$\{C_т(n)\}$, n=1,2,...N, где каждой строке приписывается номер $C_т(n)$ и N – число строк ПП.

2. Вычленение (токенизация) в строке цепочек буквенных символов, ограниченных справа словоразделом, знаком пунктуации или конечным символом и представляющих собой графическое слово или тире, с удалением символов пунктуации
$C_т=q_1q_2…q_{r(0)}s…sq_1q_2…q_{r(i)}s…sq_1q_2…q_{r(k)}$,
где q∈{а,б,в,…я}, i=1,2,...(k-1), и s – словораздел.

3. Акцентуация графических слов производится на основе «Полной акцентуированной парадигмы по А. А. Зализняку» [13], где простановка ударений на гласных буквах осуществлена посредством апострофа – q', q∈{а,о,э,и,у,ы,е,ё,ю,я}. При этом опускаются имеющиеся в словах второстепенные

ударения и необязательные ударения в словах с пометой *(часто без удар.)*. Для слов с пометой *(нормально без удар.)* в сочетании со следующим словом осуществляется поиск в списке словосочетаний, где данное слово получает, а следующее теряет свое ударение.

4. Для выделения рифмованных строк можно использовать такие ресурсы как генераторы рифм или же можно отождествлять рифмы с помощью оригинальных алгоритмов, но в этом случае требуется орфоэпическая трансформация всего текста или конечной группы слогов с последним акцентуированным словом в каждой строке ПП. Здесь же предполагается, что определение рифмованных строк в ПП может осуществляться с помощью «Большого словаря рифм» [6]. В строке, начиная с первой, для конечной группы слогов с последним акцентуированным словом $q_1 \dots q_{i-1}q_i'q_{i+1}\dots q_{r(k)}$ ищется набор рифм $\{P_ф1\}$. Последние слова последующих строк проверяются на присутствие в этом наборе. После проверки последнего слова последней строки ищется набор рифм $\{P_ф2\}$ к последнему слову следующей строки, не нашедшемуся в предыдущем наборе. И далее последние слова последующих строк проверяются на присутствие в новом наборе. Процедура повторяется до тех пор, пока не останется ни одного последнего слова, не нашедшего себе подходящего набора. Каждой строке, получившей рифму из одного из наборов, приписывается наряду с номером и код (порядковый номер) этого набора $(n,P_ф j)$, n=1,2,...N. В итоге каждой рифме (набору рифм) ПП соответствует некоторая последовательность номеров строк $P_ф j\{n\}$, и совокупность $\{(n,P_ф j)\}$ задает в спецификации схему рифмовки. Может оказаться так, что $ПП=\{C_т(n,P_ф n)\}$, т.е. текст ПП состоит из нерифмованных между собой строк, когда каждой строке соответствует свой набор рифм.

5. Выведение из графического представления стиха п.2. слоговой схемы ($C_л$) в виде цепочек гласных букв в графических словах строки, без выделения ударений и со словоразделами после удаления всех согласных:

$C_л=c_1c_2\dots c_{r(0)}s\dots sc_1c_2\dots c_{r(i)}s\dots sc_1c_2\dots c_{r(t)}$,

где $c\in\{а,о,э,и,у,ы,е,ё,ю,я\}$, i=1,2,...(t-1), и s – словораздел.

6. Представление цепочки словоразделов с номерами предшествующих слогов (гласных):

$s_{r(0)}\dots s_{r(0)+\dots+r(i)}\dots s_{r(0)+\dots+r(i)+\dots+r(t-1)}$.

Выявление среди словоразделов каждой строки словораздела с одним и тем же номером для всех строк ПП $s_{r(0)+\dots+r(i)}$, где $j=r(0)+\dots+ r(i)=const$. Таким образом, в ПП выявляется место цезуры $S_j=s_{r(0)+\dots+r(i)}$ –словораздел на постоянном месте во всех строках ПП, отсутствие цезуры обозначим как S_0.

7. Выведение слоговой схемы в виде цепочек ударных и безударных слогов без словоразделов:

$C_л=c_1\dots c_{m(0)}C_1c_1\dots c_{m(1)}\dots C_ic_1\dots c_{m(i)}\dots C_{k-1}c_1\dots c_{m(k-1)}\ C_k\ c_1\dots c_{m(k)}$,

где с – безударный слог, $(c_1\dots c_{m(i)})$ – анакруза, междуакцентный интервал и клаузула, $0\le i\le k$, C_i – ударный слог, $1\le i\le k$, и k – число ударных слогов.

Полученные характеристики строки в значительной мере специфицируют ее метрико-ритмический характер. Среди безударных слогов не исключены акцентологически неопределенные слоги, которые могут принимать на себя ударение. Снять эту неопределенность для языка с подвижным ударением может помочь сравнение с соответствующими характеристиками других строк стихотворения, и тогда можно установить какие из этих характеристик являются регулятивом, т.е. определенным сопрягающим и разделяющим признаком, повторяющимся в смежных или соотнесенных по структуре строках ПП [2].

4 Набор ритмических шаблонов стиха для идентификации стихового метра ПП

Решение данной задачи идентификации заключается в сопоставлении ритмических вариантов стиха изучаемого ПП с набором ритмических шаблонов из определенного репертуара метрико-ритмических вариантов стиха [7].

Для этого необходимо вывести схемы (формулы) метрико-ритмических вариантов стиха.

Примем следующие обозначения для некоторых стиховедческих понятий в формальном языке для метрических, метрико-ритмических и ритмических схем стиха, представляющих собой цепочку символов ударных и безударных слогов и словоразделов между тактовыми группами (ритмическими словами).

В метрической (идеальной) схеме стиха:

A – сильное место в строке, предназначенное для ударного слога и называемое зачастую иктом;

b – слабое место в строке, предназначенное для безударного слога.

В метрико-ритмической (вариантной) схеме стиха:

A – сильное место в строке с ударным слогом;

b – слабое место в строке с безударным слогом;

a – сильное место в строке с безударным слогом;

B – слабое место в строке с безударным слогом;

В ритмической (реальной) схеме стиха:

C – ударный слог в строке;

c – безударный слог в строке;

Для цепочек безударных слогов:

$c_1c_2\dots c_k=c^k$, $c^0=1$, c1=c и 1c=c.

В последующих формулах (схемах) стиха словоразделы исключены, поскольку наличие и место цезуры в стихах ПП было выяснено в п. 6. предыдущего раздела.

4.1 Ритмические схемы акцентного стиха

Исходя из принятых обозначений можно дать **слоговую ритмическую схему стиха**:

$C_p=c^{r(0)}C_1c^{r(1)}C_2c^{r(2)}\ \dots\ C_ic^{r(i)}C_{i+1}\dots\ C_{k-1}c^{r(k-1)}C_kc^{r(k)}$,

где $0 \le r(i)$ – число безударных слогов, $0 \le i \le k$,

$c^{r(i)}$ – междуакцентный интервал, $1 \le i \le (k-1)$,

$c^{r(0)}$ – анакруза (безакцентное начало стиха),

$c^{r(k)}$ – клаузула (безакцентное окончание стиха).

Ударный слог с последующим междуакцентным интервалом условно считается ритмической группой стиха, так что C_p можно свернуть в 3-частную схему:

(1) $\mathbf{C_p = c^{r(0)}(\prod_{i=1,2,..(k-1)} C_i c^{r(i)}) A b^{r(k)}}$,

где $1 \le i \le (k-1)$ и первая часть (начало стиха) представлена анакрузой, вторая часть (середина стиха) представлена цепочкой ритмических групп, третья часть (окончание стиха) представлена последним k-м обязательно ударным слогом $C_k = A$ (правило константной ударности последнего икта [15]) и клаузулой с обязательно безударными слогами $c^{r(k)} = b^{r(k)}$.

Размерность P_c строки с номером n с учетом предварительной разметки можно представить в спецификации совокупностью места клаузулы S_j, числа ударных слогов k, общего числа слогов в строке $R = (k + \sum r(i)_{i=0,1,\dots k})$ и цепочки чисел слогов анакрузы, междуакцентных интервалов и клаузулы $\{r(i)\} = (r(0), r(1), r(2), \dots r(i-1), r(i), \dots r(k-1), r(k))$, $i = 0,1,\dots(k)$:

$P_c \langle n, S_j, k, R, r(0), \{r(i)\}_{i=1,2,\dots(k-1)}, r(k) \rangle$.

С учетом пояснения к схеме (1) в этом выражении можно объединить два параметра R и r(k) в один $(R\text{-}r(k)) = (k + r(0) + \sum r(i)_{i=1,\dots(k-1)})$, в котором число ударных слогов k и число безударных слогов без клаузулы связаны, так что размерность строки характеризуется, в первую очередь, параметрами k и (R-r(k)):

(2) $P_c \langle n, S_j, \mathbf{k, (R\text{-}r(k))}, r(0), \{r(i)\}_{i=1,2,\dots(k-1)} \rangle$.

Таким образом, клаузула, как и рифма, не является одним из определяющих параметров для размерности строки и метрико-ритмической структуры стиха.

Условие 1 для параметров размерности (2).

Если для всех n=1,2,...N параметры $(R\text{-}r(k)) \neq \text{const}$ и $k \neq \text{const}$, то схема (1) может представлять два вида схем в зависимости от модуля разности Δ_{fg} между величинами параметра (R-r(k)) для разных строк ПП, f,g=1,… N. Значения $\Delta_{fg} \in \{\Delta_{fg}\}$ лежат в интервале натурального ряда (1,2,3,…20).

Условие 1М для разности Δ_{fg} между строками ПП. Если

1) все значения Δ_{fg} кратны только 2, $\Delta_{fg} \in (2,4,6,8,10,12,14,16,18,20)$;

2) все значения Δ_{fg} кратны только 3, $\Delta_{fg} \in (3,6,9,12,15,18)$;

3) все значения Δ_{fg} кратны только 5, $\Delta_{fg} \in (5,10,15,20)$,

то схема (1) с большой вероятностью представляет ритмическую схему **неурегулированного (неравностопного) метрического стиха**, или **вольного силлабо-тонического стиха**.

Условие 1Д для разности Δ_{fg} между строками ПП.

Если все значения Δ_{fg} лежат в интервале натурального ряда, $\Delta_{fg} \in (1,2,3,\dots 20)$, и не удовлетворяют исключительным условиям 1), 2) и 3), то схема (1) представляет ритмическую схему **неурегулированного акцентного стиха**, или **дисметрического стиха**.

По отношению к рифме можно выделить два вида дисметрического стиха:

(1Д.а) схема **свободного рифмованного стиха** или **говорного (раешного) стиха** при $C_т(n, P_ф j)$, $j \neq n$, $j \ge 2$, для всех n=1,2,...N;

(1Д.б) схема **свободного нерифмованного стиха** или **верлибра** при $C_т(n, P_ф n)$ для всех n=1,2,...N.

Условие 2 для параметров размерности (2).

Если для всех n=1,2,...N параметры $(R\text{-}r(k)) \neq \text{const}$ и k=const, то схему (1) можно представить в виде схемы урегулированного акцентного стиха:

(3) $\mathbf{b^{r(0)}(\prod_{i=1,2,..(k-1)} A_i b^{r(i)}) A b^{r(k)}}$,

где $b^{r(0)}$ – анакруза, $b^{r(k)}$ – клаузула, $(\prod_{i=1,2,..(k-1)} A_i b^{r(i)})$ – цепочка ритмических групп с междуакцентными интервалами, $1 \le i \le (k-1)$.

Схема (3) представляет по отношению к рифме два вида урегулированного акцентного стиха:

(3.а) схема **рифмованного k-акцентного тонического стиха** при $C_т(n, P_ф j)$, $j \neq n$, $j \ge 2$, для всех n=1,2,...N;

(3.б) схема **нерифмованного k-акцентного тонического стиха** при $C_т(n, P_ф n)$ для всех n=1,2,...N.

С другой стороны ограничения в размерности (2) величин междуакцентных интервалов подразделяет независимо от наличия или отсутствия рифмы схему (3) на два таких вида:

(3.1) схема **неурегулированного k-акцентного тонического стиха** при $r(i)_{max} \ge 4$, $1 \le i \le (k-1)$, для всех n=1,2,...N;

(3.2) схема **урегулированного k-акцентного тонического стиха** при $r(i)_{max} \le 3$, $1 \le i \le (k-1)$, для всех n=1,2,...N.

Дальнейшие ограничения на величины междуакцентных интервалов в схеме (3.2) подразделяет этот вид на следующие подвиды:

(3.2.Т4) схема урегулированного k-акцентного тонического стиха с 4-сложным диапазоном ритмической группы или **4-сложного k-акцентного тактовика** при $1 \le r(i) \le 3$, $1 \le i \le (k-1)$, для всех n=1,2,...N;

(3.2.Т3) схема урегулированного k-акцентного тонического стиха с 3-сложным диапазоном ритмической группы или **3-сложного k-акцентного тактовика** при $0 \le r(i) \le 2$, $1 \le i \le (k-1)$, для всех n=1,2,...N;

(3.2.Д3) схема урегулированного k-акцентного тонического стиха с 3-сложным диапазоном ритмической группы или **3-сложного k-акцентного дольника** при $1 \le r(i) \le 2$, $1 \le i \le (k-1)$, для всех n=1,2,...N;

(3.2.Д2) схема урегулированного k-акцентного тонического стиха с 2-сложным диапазоном ритмической группы или **2-сложного k-акцентного**

дольника при 0≤r(i)≤1, 1≤i≤(k-1), для всех n=1,2,...N.

4.2 Метрические схемы метрического стиха

Условие 3 для параметров размерности (2).

Если для всех n=1,2,...N параметры (R-r(k))≠const, k=const и r(i)=const (r(i)=r – постоянный междуиктовый интервал), то схему (3) можно представить в виде схемы **неурегулированного k-стопного силлабо-тонического стиха**:

(4) $\mathbf{b^{r(0)}(Ab^{r})^{k-1}Ab^{r(k)}}$,

где $b^{r(0)}$ – анакруза, $b^{r(k)}$ – клаузула, $(Ab^{r})^{k-1}$ – цепочка (k-1) ритмических групп с равными междуиктовыми интервалами, 1≤i≤(k-1), так что (k-1)-кратная (r+1)-сложная ритмическая группа представляет собой своеобразный аналог стопы или краты [9].

Длина слоговой строки R=(k+r(0)+r(k)+(к-1)r).

Схема (4) представляет по отношению к рифме два вида схем:

(4.а) схема **рифмованного неурегулированного (с переменной анакрузой) k-стопного силлабо-тонического стиха** при $C_т(n,P_ф j)$, j≠n, j≥2, для всех n=1,2,...N;

(4.б) схема **нерифмованного (белого) неурегулированного (с переменной анакрузой) k-стопного силлабо-тонического стиха** при $C_т(n,P_ф n)$ для всех n=1,2,...N.

Условие 4 для параметров размерности (2).

Если для всех n=1,2,...N параметры (R-r(k))=const, k=const и r(i)=const (r(i)=r – постоянный междуиктовый интервал), то r(0)=const (постоянная анакруза), и схемой (4) представляется

(5) схема **строго урегулированного k-стопного силлабо-тонического стиха.**

Схема (5) представляет по отношению к рифме два вида схем:

(5.а) схема **рифмованного строго урегулированного k-стопного силлабо-тонического стиха** при $C_т(n,P_ф j)$, j≠n, j≥2, для всех n=1,2,...N;

(5.б) схема **нерифмованного (белого) строго урегулированного k-стопного силлабо-тонического стиха** при $C_т(n,P_ф n)$ для всех n=1,2,...N.

Далее независимо от наличия или отсутствия рифмы схема (5) в зависимости от длины междуиктового интервала r и длины клаузулы r(0) для всех n=1,2,...N порождает схемы силлабо-тонических размеров стиха.

Размеры с 2-сложной стопой, r=1:

(5.2-Х) $\mathbf{(Ab)^{k-1}Ab^{r(k)}}$,

r(0)=0, – схема **к-стопного хорея;**

(5.2-Я) $\mathbf{b(Ab)^{k-1}Ab^{r(k)}}$,

r(0)=1, – схема **к-стопного ямба**.

Размеры с 3-сложной стопой, r=2:

(5.3-Д) $\mathbf{(Ab^{2})^{k-1}Ab^{r(k)}}$,

r(0)=0, – схема **к-стопного дактиля**;

(5.3-Ам) $\mathbf{b(Ab^{2})^{k-1}Ab^{r(k)}}$,

r(0)=1, – схема **к-стопного амфибрахия**;

(5.3-Ан) $\mathbf{b^{2}(Ab^{2})^{k-1}Ab^{r(k)}}$,

r(0)=2, – схема **к-стопного анапест**а.

Размеры с 4-сложной стопой, r=3:

(5.4-Пн1) $\mathbf{(Ab^{3})^{k-1}Ab^{r(k)}}$,

r(0)=0, – схема **к-стопного пэона-1**;

(5.4-Пн2) $\mathbf{b(Ab^{3})^{k-1}Ab^{r(k)}}$,

r(0)=1, – схема **к-стопного пэона-2**;

(5.4-Пн3) $\mathbf{b^{2}(Ab^{3})^{k-1}Ab^{r(k)}}$,

r(0)=2, – схема **к-стопного пэона-3**;

(5.4Пн4) $\mathbf{b^{3}(Ab^{3})^{k-1}Ab^{r(k)}}$,

r(0)=3, – схема **к-стопного пэона-4**.

Размеры с 5-сложной стопой, r=4:

(5.5-Пт1) $(\mathbf{Ab^{4}})^{k-1}Ab^{r(k)}$,

r(0)=0, – схема **к-стопного пентона-1**;

(5.5-Пт2) $\mathbf{b(Ab^{4})^{k-1}}Ab^{r(k)}$,

r(0)=1, – схема **к-стопного пентона-2**;

(5.5-Пт3) $\mathbf{b^{2}(Ab^{4})^{k-1}}Ab^{r(k)}$,

r(0)=2, – схема **к-стопного пентона-3**;

(5.5-Пт4) $\mathbf{b^{3}(Ab^{4})^{k-1}}Ab^{r(k)}$,

r(0)=3, – схема **к-стопного пентона-4**;

(5.5-Пт5) $\mathbf{b^{4}(Ab^{4})^{k-1}}Ab^{r(k)}$,

r(0)=4, – схема **к-стопного пентона-5**.

Такие экзотические размеры как **логаэды** здесь из рассмотрения опускаются.

Условие 5 для параметров размерности (2).

Если для всех n=1,2,...N параметры (R-r(k))=const и k≠const, то r(i)≠const, 1≤i≤(k-1), и схемой (3) по отношению к рифме могут представляться два вида схем **урегулированного изосиллабического стиха:**

6.а) схема **рифмованного урегулированного изосиллабического стиха** при $C_т(n,P_ф j)$, j≠n, j≥2, для всех n=1,2,...N;

(6.б) схема **нерифмованного (белого) урегулированного изосиллабического стиха** при Ст(n,Рфn) для всех n=1,2,...N.

При Условии 5 по причастности к стопному стиху из схемы (3) выделяется

(6.1) схема **урегулированного силлабо-тонического стиха** с различными нарушениями.

Во-первых, это пропуск схемных ударений:

пиррихий в размерах с 2-сложной стопой (сс);

трибрахий в размерах с 3-сложной стопой (ссс).

Во-вторых, это наложение сверхсхемных ударений:

спондей в размерах с 2-сложной стопой;

в размерах с 3-сложной стопой – тримакр (ССС), амфимакр (СсС) для дактиля (Ссс) и анапеста (ссС), антибакхий (ССс) для дактиля (Ссс) и амфибрахия (сСс), бакхий (сСС) для амфибрахия (сСс) и анапеста (ссС).

Число метрико-ритмических вариантов схемы (6.1) трудно поддается подсчету и в настоящем исследовании не рассматривается.

При непричастности к стопному стиху и добавлении к Условию 5 требование r(k)=1 из схемы (3) выделяется схема **R-сложного силлабического стиха**

(6.2) $\mathbf{b^{r(0)}(\prod_{i=1,2,..(k-1)}A_i b^{r(i)})Ab}$.

Из схемы (6.2) можно выделить две схемы по отношению к цезуре:

(6.2) схема **R-сложного бесцезурного силлабического стиха** для R=7,8,9,10;

(6.2) схема **R-сложного цезурированного силлабического стиха** для R=11,12,13,14,15,16, **е**сли имеется цезура S_j, j≥5.

4.3 Пример идентификации ямбического размера

Рассмотрим метр стихотворения Ф.И. Тютчева:

Как дымный столп светлеет в вышине! –
Как тень внизу скользит неуловима!..
«Вот наша жизнь – промолвила ты мне, –
Не светлый дым, блестящий при луне,
А эта тень, бегущая от дыма…»

Метрико-ритмическая схема этого ПП:

B Ab A bAb abA
B A bA bA babAb
B Ab A bAba B A
b Ab A bAb a bA
b Ab A bAba b Ab

Поскольку метр ПП заранее не определен, то автоматический анализатор метра будет иметь дело с ритмической схемой без учета словоразделов:

CCcCcCcccC
CCcCcCcccCc
CCcCcCccCC
cCcCcCcccC
cCcCcCcccCc

В схеме словоразделов имеются два проходящих через все строки ПП словораздела – после первого слога и после 4-го, последний как раз является мужской цезурой S_4, поскольку следует сразу после ударного слога и делит каждый стих на два полустишия:

1s 23s **4s** 567s 8910
1s 2s 3**4s** 56s 7891011
1s 23s **4s** 5678s 9s 10
1s 23s **4s** 567s 8s 910
1s 23s **4s** 5678s 9s 1011

Нетрудно задать нумерацию строк и определить схему рифмовки, затем вычислить длину анакруз, междуакцетных интервалов и клаузул с тем, чтобы выписать схему размерностей строк:

(1,$P_ф$1), 0C0C1C$S_4$1C3C0, k=5, r(0)=0, 0≤ r(i)≤3, r(k)=0, R=10

(2,$P_ф$2), 0C0C1C$S_4$1C3C1, k=5, r(0)=0, 0≤ r(i)≤3, r(k)=1, R=11

(3,$P_ф$1), 0C0C1C$S_4$1C2CC0, k=6, r(0)=0, 0≤ r(i)≤2, r(k)=0, R=10

(4,$P_ф$1), 1C1C$S_4$1C3C3C0, k=4, r(0)=1, 1≤ r(i)≤3, r(k)=0, R=10

(5,$P_ф$2), 1C1C$S_4$1C3C3C1, k=4, r(0)=1, 1≤ r(i)≤3, r(k)=1, R=11

Таким образом, обобщенная размерность ПП такова:

((1,3,4, $P_ф$1), (2,5, $P_ф$2); S_4; k=5,5,6,4,4; (R-r(k)=10; r(0)=0,0,0,1,1; 0≤r(i)≤3).

Далее, обращаясь к Условиям раздела 4, можно найти, что выявленные размерности ПП удовлетворяют только Условиям 5, откуда следует, что рассматриваемое ПП соответствует метрическим схемам (6а) и (6.1) силлабо-тонических размеров, а не схеме (6.2) силлабических, поскольку r(k)=0 в 3-х случаях из 5.

Поскольку соотношение числа ударных слогов в строке и длины строки удовлетворяют только двусложному размеру (5.2), а наличие женских анакруз исключает размер хорея, то наиболее подходящей является схема (5.2Я) **k-стопного ямба**, и легко вычислить, что k=5.

Накладывая на ритмическую схему ПП метрическую схему 5-стопного ямба, можно получить метрико-ритмическую схему ПП в первом приближении:

CAcAcAcacA
CAcAcAcacAc
CAcAcAcaCA
cAcAcAcacA
cAcAcAcacAc

На этой схеме сразу же обнаруживаются пиррихии **a** в последней ритмической группе каждой строки. Чтобы получить искомую метрико-ритмическую схему, остается только обозначить спондеи B в первых трех строках, сделав замену C на B, и, заменив **c** на **b**, обозначить метрически безударные слоги.

5 Программно-технические решения по реализации процедур спецификации ПП

В качестве инструмента для реализации применим высокоуровневый язык программирования Python, который подходит для обработки русского языка, и совместим с фреймворком Django для дальнейшего использования единого разработанного алгоритма на web-ресурсе.

Программная разработка спецификации ПП условно подразделяется на 3 задачи:

1. Подготовка данных;
2. Метрико-ритмическая разметка;
3. Идентификация стиха.

В связи с тем, что возможна неопределенность при простановке ударений, поиске рифмы и выборе метрической схемы, то предусмотрено в случае сомнительной идентификации метра обращение к экспертным решениям в части метрико-ритмической разметки текста ПП, а также использование методов машинного обучения для классификации неоднозначных ситуаций.

6 Заключение

В работе представлены методы и процедуры, которые позволяют автоматически осуществлять спецификацию ПП по основным размерностям строк и произведения в целом. Выработанный подход в основном позволяет обходить трудности акцентологической разметки и снятия омонимии

при идентификации метра ПП. Намеченные процедуры обещают быть продуктивными для анализа ПП с помощью такого сетевого ресурса как ИАСРП.

Литература

[1] Бойков В.Н., Захаров В.Е., Пильщиков И.А., Сысоев Т.М. Тезаурус как инструмент поэтологии // Моделирование и анализ информационных систем. 2010. Т. 17, № 1. С. 5–24. *(Boykov V.N., Zakharov V.E., Pilshchikov I.A., Sysoev T.M. Thesaurus as a Poetological Tool // Modeling and analysis of information systems. 2010. V. 17, No 1. P. 5–24 [in Russian])*.

[2] Бойков В.Н. Контекстно-свободная грамматика одной ритмической модели русского стиха. // Моделирование и анализ информационных систем. – 2012. – Т. 19, № 4. – С. 154

[3] Бойков В.Н., Захаров В.Е., Каряева М.С., Соколов В.А. Тезаурус по поэтологии как инструмент для информационного поиска и коллекции знаний. // Моделирование и анализ информационных систем. – 2013. – Т. 20, № 4. – С. 5–24.

[4] Бойков В.Н., Пильщиков И.А. Семантическая модель «Тезауруса по поэтологии» в составе информационно-аналитической системы // Интернет и современное общество: сборник научных статей. Труды XVI Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2013), Санкт-Петербург, 9 – 11 октября 2013 г. — СПб.: НИУ ИТМО, 2013. – С. 273–278.

[5] Бойков В.Н., Захаров В.Е., Каряева М.С., Соколов В.А. Об автоматической рубрикации терминов тезауруса открытой информационно-аналитической системы. // Электронные библиотеки: перспективные методы и технологии, электронные коллекции, RCDL - Дубна, 2014.

[6] Большой словарь рифм http://rifmovnik.ru/docs.htm

[7] Гаспаров М.Л. Очерк истории русского стиха. Метрика. Ритмика. Рифма. Строфика. – М.: Фортуна Лимитед, 2-е издание, 2002.

[8] Зализняк А.А. Грамматический словарь русского языка. Словоизменение. – Москва: Русский язык, 1980.

[9] Квятковский А.П. Поэтический словарь. – М.: Советская энциклопедия, 1966. (http://feb-web.ru/feb/kle/default.asp?/feb/kle/kle.html)

[10] Крылов С.А., Старостин С.А. 2003 — Актуальные задачи морфологического анализа и синтеза в интегрированной информационной среде STARLING // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2003» (Протвино, 11-16 июня 2003 г.). - М., 2003. - С. 354-360.

[11] Открытый грамматический словарь русского языка. http://odict.ru

[12] Пильщиков И.А., Старостин А.С. Автоматическое распознавание метра: проблемы и решения. // Славянский стих. IX. – М.: Рукописные памятник и Древней Руси, 2012. – С. 492–498

[13] Полная акцентуированная парадигма по А. А. Зализняку. Идея и программное обеспечение Андрея Усачёва, 2004. http://www.speakrus.ru/dict/

[14] Старостин С.А. 1994 — Рабочая среда для лингвиста // Гуманитарные науки и новые информационные технологии. - 1994. - № 2. - С. 7-22.

[15] Jakobson R. Closing Statement: Linguistics and Poetics // Style in Language / Ed. by T. A. Sebeok. - New York - London. - 1960. - P. 350-377.

[16] Hayward M. A connectionist model of poetic meter //Poetics. – 1991. – 20(4), 303-317.

[17] Rodolfo D., CIPRIAN B. SPARSAR: a System for Poetry Automatic Rhythm and Style AnalyzeR //SLATE 2013. – Grenoble University, 2013. – C. 95-95.

[18] SPARSAR https://sparsar.wordpress.com/

[19] WordNet https://wordnet.princeton.edu/

On an Automatic Procedure for the Specification of a Poetic Text for an Open Information-Analytical System

V.N. Boikov, M.S. Karyaeva, V.A. Sokolov, I.A. Pilshchikov

The paper deals with such automatic procedure for the specification of a poetic text as the metric-rhythmic marking and the identification of verse meters. Procedure algorithmization for identification of verse dimensions of such a meter as "iambic" looks promising to develop approaches for the entire spectrum of Russian versification systems.

The work was supported by the Russian Foundation for Basic Research, grant № 13-06-00448.