Об автоматизации комплексного анализа русского поэтического текста

© В. Б. Барахнин © О. Ю. Кожемякина
Институт вычислительных технологий СО РАН,
Новосибирский государственный университет,
Новосибирск
bar@ict.nsc.ru olgakozhemyakina@mail.ru

Аннотация

Целью настоящей работы является выработка подходов и технологий для автоматизации комплексного анализа русского поэтического текста. Результаты такого анализа позволят существенно расширить возможности филологов, в том числе уменьшить зависимость качества сравнительного анализа от личной эрудиции исследователя, а также применять различные методы интеллектуального анализа данных.

Работа выполнена при частичной поддержке РФФИ (проекты 10-07-00302, 11-07-00561, 12-07-00472), президентской программы «Ведущие научные школы РФ» (грант НШ 6293.2012.9) и интеграционных проектов СО РАН.

1 Постановка задачи

Современный подход к исследованию текстовых сообщений предполагает использование многоуровневой модели информации, изложенной, например, в работе германского исследователя В. Гитта [1]. Структура модели представлена на рис. 1.

Анализируя эту модель, нетрудно видеть, что ее нижний уровень соответствует шенноновскому значению термина «информация», три последующих – семиотической триаде (синтактика – семантика – прагматика), а верхний уровень носит, скорее, философский характер. При этом наличие в некотором сообщении информации высокого уровня влечет за собой наличие информации всех низших высоких уровней, но, разумеется, не наоборот (еще раз напомним: объем информации зависит, в том числе, от характеристик адресата, причем это касается всех уровней информации).

Вполне очевидно, что 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, например, к сообщениям (текстам) художественного жанра, прежде всего, – к поэтическим текстам. Достаточно вспомнить известную книгу Ю.М. Лотмана [6], в которой утверждается, что «явление структуры в стихе всегда в конечном итоге оказывается явлением смысла».

Уровни структуры стиха, подобно уровням структуры произвольного сообщения, также представляют собой определенную иерархию (см., например, [8]): являются *метр, ритм, фонетика, лексика, грамматика, речевой жанр (композиционно-речвое целое), тематика, литературный жанр*. При этом процесс анализа стиха предусматривает первоначальное рассмотрение каждого уровня как самостоятельной смысловой единицы с последующим связыванием этих наблюдений с другими элементами структуры.

Нетрудно заметить, что между уровнями структуры произвольного сообщения и стиха наблюдается определенная корреляция: к синтаксическому уровню соответствуют метр, ритм и фонетика (согласно В.Гитту, система символов сообщения относится к именно синтаксическому уровню информации), к семантическому – лексика и грамматика. Что же касается тематики, то применительно к анализу стихов она относится не только (и во многом даже не столько) к семантическому, но и прагматическому уровню, поскольку при анализе лирического стихотворения анализ тематики нередко включает исследование эмоционального *воздействия* на читателя. Наконец анализ жанров (речевого и литературного предполагает исследование сообщений *внутри стихотворного текста*: ибо, согласно в [8], речевой жанр подразумевает не только определенный тип речевого субъекта, но и столь же определенный тип речевого адресата, взаимодействие речевого субъекта и речевого адресата и создает специфику того или иного литературного жанра. На данном этапе исследования задачи

Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.

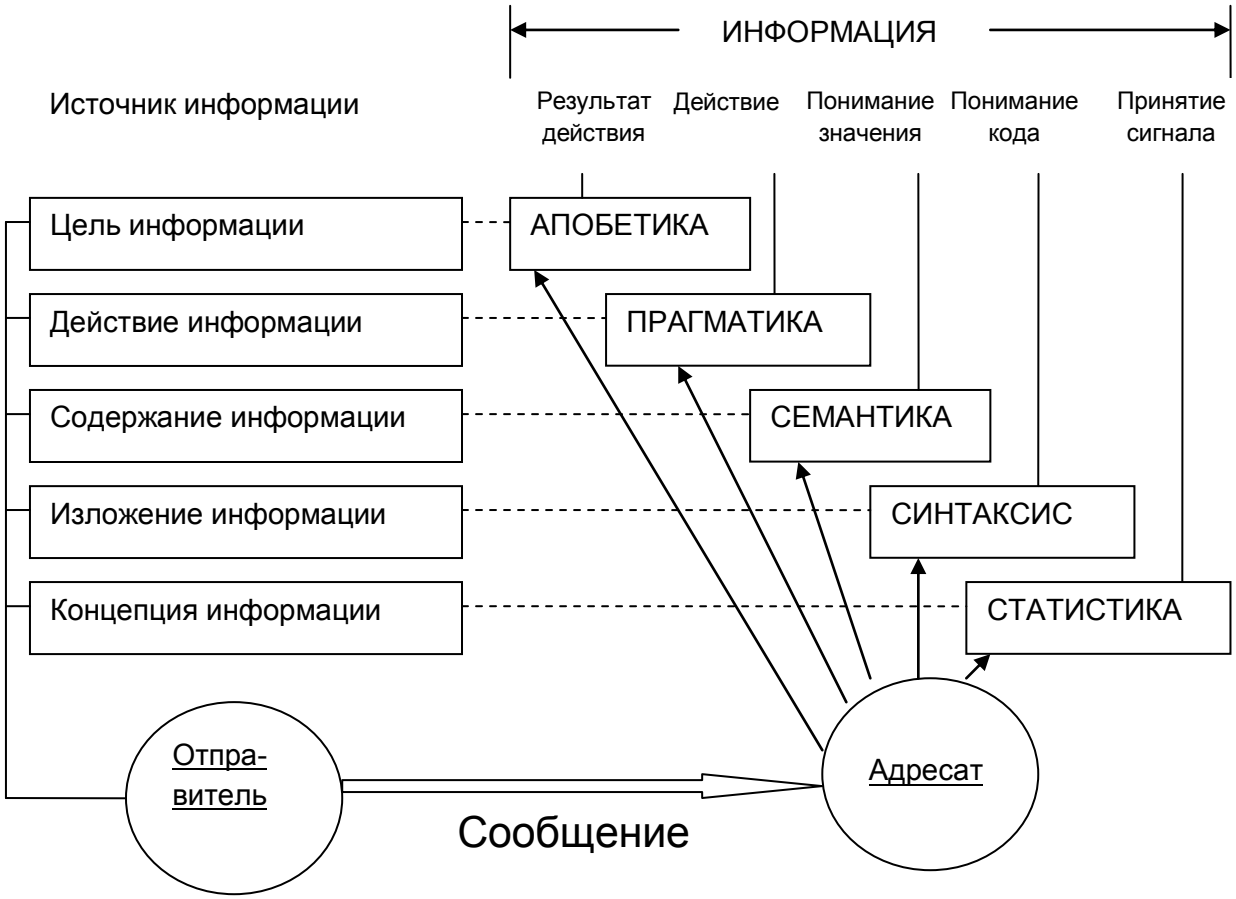

Рис. 1. Пятиуровневая модель информации В.Гитта.

жанрового анализа не рассматриваются.

Хотя отдельные работы в области исследования влияния нижних уровней структуры стиха на высшие появились еще в первой половине XX века (например, в книге К.И.Чуковского [16] среди прочего обсуждается вопрос о влиянии гласных звуков в стихах А.А.Блока на их эмоциональные характеристики), систематическое изучение такого влияния началось, по-видимому, с работ американского филолога К. Тарановского, выступившего в 1963 году на Пятом съезде славистов с докладом «О взаимодействии стихотворного ритма и тематики», в котором на основе анализа нескольких десятков русских стихотворных текстов было исследовано взаимодействие ритмических особенностей и жанрового применения пятистопного хорея. Было показано, что во многих стихах, написанных этим размером (начиная с «Выхожу один я на дорогу... М.Ю.Лермонтова), «динамический мотив пути противопоставляется статическому мотиву жизни» (см. [15]). В указанной работе была предложена методика определения семантики того или иного стихотворного размера, заключающаяся в исследовании не его единичные употребления, а по традицию его жанрового и тематического использования [8], что предполагает анализ корпусов поэтических текстов.

Систематические исследования в этом направлении были продолжены М. Л. Гаспаровым, который, в частности, показал [4], что «число метров в стиховой культуре обычно бывает сравнительно невелико, число типичных построений содержания – во много раз больше, поэтому один и тот же метр может служить знаком нескольких и даже многих тематических рядов. <...> В таких случаях, когда мы приступаем к стихотворению, то, воспринимая метр, угадываем сразу некоторый набор обычных в нем тематических ожиданий, а воспринимая лексику, устанавливаем, какой вариант из этого набора избран автором. <...> Лексика формирует для нас прежде всего семантику данного конкретного стихотворения, метрика – общий фон семантической традиции, на котором оно воспринимается».

Итак, исследование исследования влияния нижних уровней структуры стиха на высшие является весьма актуальной проблемой русской филологии. Одной из основных трудностей при ее решении является необходимость анализа корпусов поэтических текстов большого объема. Задача эта чрезвычайно трудоемкая, поэтому зачастую в поле зрения исследователя попадает лишь сравнительно небольшой круг стихов поэтов-классиков, что, без сомнения, значительно снижает полноту анализируемого материала и,

следовательно, достоверность полученных результатов. Таким образом, возникает задача автоматизации анализа различных уровней структуры стиха, что позволит освободить исследователей от рутинной работы и при этом резко расширить круг анализируемых авторов.

Описанная выше корреляция между уровнями структуры произвольного сообщения и стиха показывает, что многие технологии и математические методы, применяемые в информатике, вполне могут быть использованы в процессе автоматизации анализа стихов.

Разумеется, простейшие математические подходы применяются в филологическом анализе русских стихотворений довольно давно. Широко известны частотные словари языка поэтов-классиков. Проводились многочисленные исследования статистики типов русской рифмы (в том числе, и применительно ко временно́й динамике), обобщенные в [11]. Однако зачастую сбор статистической информации до сих пор ведется практически вручную (исключение составляет лишь контент-анализ). Отдельные исследования, описывающие комплексный подход к автоматизации характеристик русских поэтических текстов (см. например, [9]), затрагивают, как правило, весьма специфические жанры поэзии – например, фольклорные стихи, структурные характеристики которых, например метрика, тематика и т.д., значительно отличаются от соответствующих структур «литературного» стиха. Отметим, что исследования зарубежных авторов в рассматриваемой области нам неизвестны.

Целью настоящей статьи является выработка подходов и технологий для автоматизации комплексного анализа русского поэтического текста.

2 Подходы и технологии автоматизации анализа поэтических текстов

2.1 Метр, ритм, фонетика

Анализ данного уровня стихов имеет весьма специфический характер, поскольку требует исследования фонетических характеристик лексем, каковое при анализе обычных сообщений почти никогда не проводится.

Сразу ответим на естественный вопрос: поскольку непосредственно в письменном сообщении его фонетические характеристики отсутствуют, можно ли отнести их к изложенной выше семиотической модели? Действительно, воспринять фонетические характеристики текста может лишь адресат информации: человек или запрограммированная на решение такой задачи информационная система, но ведь то же самое можно сказать и про семантические характеристики текста, например, смысл лексем. Здесь следует руководствоваться известным утверждением А. А. Ляпунова [7]: «информация всегда относительна, она зависит от того, какой информационной системой она воспринимается», на основании которого фонетические характеристики текста вполне могут быть отнесены к его синтаксическому уровню.

Анализ метра и ритма предполагает исследование чередования так называемых сильных и слабых звуков (несколько упрощенно – ударных и безударных слогов), при этом метр – «идеальная схема» чередования, а ритм – их реальное чередование, несколько отличающееся от идеального ввиду взаимодействия естественных свойств речевого материала и метрического закона [8].

Для такого анализа используются фонетические словари. Наиболее полным из известных нам сетевых фонетических словарей открытого доступа – «Словарь полного фонетического разбора» [12].

Однако использование этого словаря для анализа фонетически характеристик стиха осложняется тем, что в нем приведены только начальные формы слов, поэтому необходима генерация фонетической записи словоформ (сами словоформы содержатся в том или ином морфологическом словаре, например, сопровождающем свободно распространяемый продукт Ispell [13]). 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.

При автоматическом анализе метра и ритма следует учитывать возможность использовать поэтом «нестандартных» ударений. Такая ситуация выявляется апостериори, посредством сравнения соответствующей строки (использование в которой «правильного» ударения нарушает общий ритм) с соседними строками.

Фонетический анализ стиха включает исследование звуковых повторов и рифм (их типов, а также строфического строения стиха, составление словарей рифм и т.п.). Поскольку историческое развитие русской рифмы характеризуется снижением ее точности, постольку при автоматизированном анализе рифмы необходимо учитывать свойства фонем. Так, согласные фонемы различаются по месту образования, по способу образования по участию голоса и шума, по твердости и мягкости, по глухоте и звонкости (подробнее см., например, [11]). Некоторые из этих свойств для каждой фонемы каждого слова непосредственно указаны в словаре [12].

Разумеется, для анализа метрических и строфических характеристик стиха необходимы «эталонные» базы даны типичных размеров и строф.

2.2 Лексика и грамматика

Лексический анализ стихотворения предусматривает [8] создание его лексического словаря, который используется, в частности, для выявления доминирующих частей речи, тематических (семантических) полей и поэтической фразеологии (прежде всего, употребляемых метафор).

Среди некоммерческих программных продуктов, решающих задачу составления лексического словаря некоторого текста, можно назвать стеммер компании «Яндекс» [14]. Он позволяет извлекать как слова, являющиеся заданной частью речи (что автоматически решает задачу выявления доминирующих частей речи), так и словосочетания заданной структуры (например, (*прилагательное*) + (*существительное*) или (*существительное*) + (*существительное в родительном падеже*)}. Последняя из названных возможностей способна значительно обогатить традиционные словари языка того или иного поэта.

Что же касается задач выявления тематических полей и метафор, то, хотя для их решения необходим лексический словарь слов и словосочетаний, они требуют и дополнительной, зачастую плохо формализуемой информации (например, о принадлежности лексем к тому или иному тематическому полю, семантическому архетипу и т.п.), и поэтому на данном этапе работы эти задачи не рассматриваются.

Грамматический анализ текста включает определение его возможной принадлежности к именному или к глагольному стилям (соответственно сплошные назывные предложения или перечисление действий), а также временно́го плана и субъектной структуры стихотворения (что требует исследования употребления категорий времени, залога и лица).

Именной или глагольный стиль определяется путем непосредственного анализа лексического словаря. Для определения употребления категорий времени, залога и лица дополнительно требуется использовать довольно несложные морфологические правила русского языка, позволяющие установить, какая конкретно категория времени, залога или лица употреблена.

2.3 Тематика

Непосредственное определение тематики стихотворения – задача, весьма сложная для автоматизированного решения, поскольку требует семантического анализа текстов на уровне, близком к восприятию естественно-языковых текстов человеком. Однако исследование зависимости тематики от низших уровней структуры стиха – одна из наименее исследованных областей филологического анализа. В этой области имеется целый ряд нерешенных проблем, некоторые из них сформулированы в [8]:

«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».

Применение методов статистического анализа больших массивов стихотворных текстов вполне может стать эффективным методом разрешения этих и подобных проблем филологического анализа.

Важным направлением исследований представляется использование многофакторного анализа семантических, эмоциональных и т.п. ассоциаций, масштабное применение которого практически невозможно без применения методов автоматизации.

Приведем пример эффективности многофакторного анализа при установлении зависимости тематической окраски произведения от его размера. В [8] для иллюстрации неоднозначности такой зависимости приводится следующий пример: «Если, скажем, рассматривается стихотворение А.С.Пушкина «Бесы», то звучание четырехстопного хорея характеризуется как «зловещее», а то и «заунывное», если же «Мойдодыр» К.Чуковского – тот же размер становится «бодрым», «стремительным», «динамичным», «игривым»». Однако, если учесть сделанное в [3] наблюдение о четырехстопном хорее, «одной из семантических окрасок которого в русской поэзии является мотив бессонницы, *утраты* [курсив наш – *авт.*] и смерти», и вспомнить начало «Мойдодыра»:

Одеяло убежало,
Улетела простыня,
И подушка, как лягушка,
Ускакала от меня

...

Боже, Боже, что случилось?
Отчего же всё кругом
Завертелось, закружилось
И помчалось колесом?

носящее, если представить описанную сцену происходящей в действительности, вполне инфернальный характер, а также учесть несомненную близость ряда семантических полей (например, связанных с быстрым беспорядочным движением) обсуждаемых произведений, то уместнее будет говорить, скорее, не о противопоставлении, а о сходстве задаваемых четырехстопным хореем семантических окрасок «Бесов» и «Мойдодыра».

Конечно, приведенный пример имеет «частный» характер. При работе с большими корпусами текстов целесообразно применение методов интеллектуального анализа данных, в частности, кластеризации. Современные подходы к кластеризации текстовых документов с использованием нескольких шкал сходства изложены, например, в монографии [17].

2.4 Об электронных библиотеках поэзии

Наконец, скажем несколько слов об электронных библиотеках поэтических текстов, которые могут послужить первичным материалом для изложенных выше исследований. Большие подборки русской поэзии, прежде всего, классической, имеются в Библиотеке Максима Мошкова [2], Интернет-библиотеке Алексея Комарова [5], на сайте «Мировое искусство: живопись, литература, анимация, кино» [10]. При этом, разумеется, при использовании этих библиотек для анализа классической поэзии могут возникнуть определенные проблемы, связанные, например, с тем, что все тесты в них приведены в современной орфографии, что способно внести известные (хотя и весьма незначительные) искажения в фонетический анализ текста.

3 Заключение

В настоящей работе намечены основные подходы к автоматизации процесса статистического анализа низших структурных уровней (метр, ритм, фонетика, лексика, грамматика) русских поэтических текстов. Результаты такого анализа позволят существенно расширить возможности филологов, исследующих как указанные уровни стихов, так и их семантические и прагматические характеристики, в том числе избавить филологов от рутинной работы, расширить круг анализируемых произведений, уменьшив зависимость качества сравнительного анализа от личной эрудиции исследователя, а также применять различные методы интеллектуального анализа данных.

Литература

[1] W. Gitt. Ordnung und Information in Technik und Natur // In: Gitt W. (Hrsg.): Am Anfang war die Information. Graefeling: Resch KG, 1982. – S. 171-211.

[2] Библиотека Максима Мошкова. http://lib.ru

[3] Винни Пух и философия обыденного языка. М: Гнозис, 2010.

[4] М. Л. Гаспаров. Семантический ореол метра: К семантике русского трехстопного ямба // В: Лингвистика и поэтика. М.: Наука, 1979. С. 282-308.

[5] Интернет-библиотека Алексея Комарова. http:///library.ru

[6] .Ю. М. Лотман. Структура художественного текста. М.: Искусство, 1970.

[7] А. А. Ляпунов. О соотношении понятий материя, энергия и информация // А.А.Ляпунов. Проблемы теоретической и прикладной кибернетики. Новосибирск: Наука, 1980. С. 320-323.

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

[9] Н. Д. Москин. Теоретико-графовые модели структуры фольклорных текстов, алгоритмы поиска закономерностей и их программная реализация // Дис. ... кандидата технич. наук. Петрозаводск, 2006.

[10] Сайт «Мировое искусство: живопись, литература, анимация, кино». http://www.world-art.ru.

[11] Д. С. Самойлов. Книга о русской рифме. М.:Художественная литература, 1982.

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

[13] Словарь русского языка для Ispell.. http://semiconductors.phys.msu.su/~swan/orthography.html

[14] Стеммер компании «Яндекс». http://company.yandex.ru/technology/mystem/

[15] К. Тарановский. О взаимоотношении стихотворного ритма и тематики // Тарановский К. О поэзии и поэтике. М.: Языки Русской культуры, 2000. С. 372-403.

[16] К. Чуковский. Александр Блок как человек и поэт. Пг.: А.Ф.Маркс, 1924.

[17] Ю. И. Шокин, А. М. Федотов, В. Б. Барахнин. Проблемы поиска информации. Новосибирск: Наука, 2010.

About the automation of the complex analysis of Russian poetic text

Vladimir Barakhnin, Olga Kozhemyakina

The purpose of this work is the development of approaches and technologies for automation of the complex analysis of the Russian poetic text. The results of such analysis will allow to expand the possibilities of philologists, and also to reduce the dependence of quality of the comparative analysis from the personal erudition of the researcher, and to apply various methods of the intellectual analysis of data.