Семантическая модель «Тезауруса по поэтологии» в составе информационно-аналитической системы♣

В. Н. Бойков, И. А. Пильщиков

Московский государственный университет имени М.В.Ломоносова
boykov_bh@bk.ru, pilshch@yandex.ru

Аннотация

Описывается логико-семантическая модель «Тезауруса по поэтологии», являющегося определенной предзаданной онтологией и системообразующим ядром проектируемой информационно-аналитической системы русской поэзии. Модель соответствует тому, что данный тезаурус создается как открытый сетевой ресурс, и призвана способствовать автоматической рубрикации тезауруса и отвечать требованиям аналитической обработки поэтических текстов.

1. «Тезаурус по поэтологии» в структуре информационно-аналитической системы русской поэзии

Проектируемая информационно-аналитическая система русской поэзии (ИАС РП) состоит предположительно из двух частей (уровней) и при этом предусматривает поочередное их создание.

Верхний уровень (1-я очередь) содержит БД «Тезаурус по поэтологии», обеспечивающую отслеживаемое модератором метаописание такой предметной области как поэтология, и представляет собой формализованную систему понятий из области теории и истории стиха, теоретической и исторической поэтики, между которыми установлены определенные семантические отношения, т.е. все необходимое для всестороннего изучения стиха и поэзии. Там же содержится пользовательский комплекс в составе интерфейса и информационно-поискового блока, обеспечивающие открытость системы для получения информации и внесения изменений и дополнений в тезаурус.

Нижний уровень (2-я очередь) содержит аналитический блок с программно-алгоритмическим комплексом, обеспечивающие спецификацию — выявление (распознавание) характерных (специфицирующих) признаков анализируемого произведения поэзии, БД поэтических произведений «Русская поэзия» и библиотеку результатов исследований (публикаций) по поэтологии, которые также могут быть объектами исследований.

Из представленной на Рис. 1 структуры ИАС РП можно видеть, что «Тезаурус по поэтологии» предназначен не только для метаописания предметной области и информационного поиска [6], но главным образом для автоматизированного анализа и спецификации поэтических, а также поэтологических текстов.

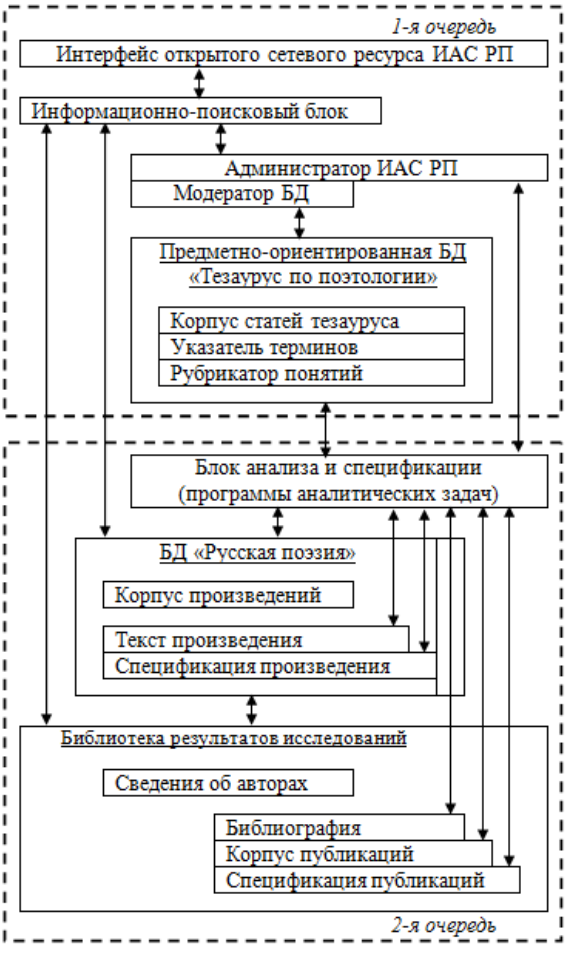

Рис. 1. Структура ИАС РП

Наметки концептуальной модели «Тезауруса по поэтологии» и его роль в качестве системообразующего ядра ИАС РП представлены в работах [1] и [5].

Интернет и современное общество: сборник научных статей XVI Всероссийской объединенной конференции IMS-2013, Санкт-Петербург, 9 - 11 октября 2013 г.

Если учесть, что объем терминологического словника тезауруса на данное время уже насчитывает около 3000 терминов, то становится понятно насколько сложным и трудоемким является процесс создания полного корпуса статей по всем терминам тезауруса. Ввиду этого понятна целесообразность решения этой задачи широким кругом специалистов и, следовательно, создание «Тезауруса по поэтологии» как открытого сетевого ресурса.

В результате сравнения редакторов онтологий, целесообразным оказывается выбор средства с использованием Wiki-технологий (Викисловарь) в качестве прототипа системы коллективного редактирования тезауруса.

Под Wiki-технологиями понимается набор средств построения веб-сайта, позволяющий пользователям самим через веб-интерфейс активно включиться в процесс редактирования его контента: исправления ошибок, добавления новых материалов и т.д. Использование Wiki-технологий для специалистов других областей не требует использования специальных программ, регистрации на сервере и знания языка HTML.

Пилотная версия информационно-логической модели «Тезауруса по поэтологии» размещена по адресу www.wiki-poetics.tw1.ru/, и основные требования по ее созданию были согласованы со стандартами по созданию тезаурусов [2], [3], [4], [15] и [16].

2. Структура логико-семантической модели «Тезауруса по поэтологии»

Нацеленность ИАС РП на аналитическую обработку текстов и ее развертывание в качестве сетевого ресурса с использованием Wiki-технологий предъявляют особые требования к логико-семантической модели «Тезауруса по поэтологии».

При разработке модели учитывался наряду с недочетами положительный опыт построения тематически близких тезаурусов [7], [8], [9], [10], [11], [12] и [13]. Можно отметить, в частности, разработки по метрическому и жанровому описанию поэтического подкорпуса Национального корпуса русского языка (www.ruscorpora.ru) на основе единой системы параметров.

Термины тезаврируемой предметной области – поэтологии — классифицируются по десяти рубрикам верхнего уровня, представляющим подобласти знания, соответствующие дисциплинам поэтологии:

1. Стиховедение;
    1.1. Стих;
        1.1.1. Метрика;
        1.1.2. Явления начала и конца стиха (строки);
        1.1.3. Ритмика;
        1.1.4. Строфика;
        1.1.5. Рифмика;
        1.1.6. Лингвистика стиха;
    1.2. Проза (в отличие от стиха);
2. Стилистика;
3. Поэтика;
4. Риторика;
5. История литературы;
6. Переводоведение и литературная компаративистика;
7. Текстология;
8. Герменевтика;
9. Теоретические школы и направления;
10. Логика и методология науки.

Здесь рубрика «Стиховедение», как наиболее формализованная дисциплина поэтологии, представлена подрубриками 2-го и 3-го уровней.

Полная рубрикация «Тезауруса по поэтологии» представляет собой определенную онтологию данной предметной области, т.е. систему понятий, отражаемую в структурированном перечне слов и устойчивых словосочетаний поэтологии (терминов/концептов/дескрипторов), в котором содержится классификация этих понятий.

Каждая терминологическая статья тезауруса (ТСТ) представляет собой комплекс семантических полей, представляющих собой две группы. Первая группа представляет различные атрибуты заглавного поля (понятия) «термин». Вторая группа представляет различные, в т.ч. отражаемые в рубрикации, отношения данного «термина» с другими терминами тезауруса.

Структура ТСТ в общем определяет логико-семантическую модель «Тезауруса по поэтологии».

Первая группа семантических полей делится в свою очередь на две части: одни из них, существенные для аналитических задач, находятся в разной степени отношения эквивалентности с заглавным термином и содержат различные дополнения к его определению, другие представляют собой разного рода отсылки и комментарии.

Вторая группа семантических полей также делится по двум типам отношений между понятиями: иерархическим (род/виды, целое/части, выше/ниже), определяющим рубрикацию тезауруса, и неиерархическим (соподчинение, смежность, ассоциация, комбинативность), устанавливающим перекрестные связи.

Всего же в ТСТ позиционируется 27 полей, и этот список открыт: в ходе постановки и решения аналитических задач может понадобиться позиционирование новых атрибутов и отношений.

3. Поля ТСТ, связанные с определением термина

1. «Термин». Слово или словосочетание на русском языке или на иностранном языке – в том случае, если в качестве единственного или основного варианта термина принят иноязычный вариант.

2. «Варианты написания» (для терминов с неустойчивой орфографией или несколькими

традициями употребления). Примеры (связей между полями):
I. 1. гекзаметр,
2. гексаметр;
II. 1. пэон,
2.1. пеон, 2.2. пеан;
III. 1. силлабо-тоника,
2. силлаботоника.

3. «Иноязычные эквиваленты». Сюда попадают эквиваленты тезавриуемого термина на иностранных языках (указываются), в которых существует устойчивая традиция его употребления и литература по темам, предполагающим использование данного термина. Оригинал иноязычного по происхождению термина (обязательно). Факультативно: перевод русского термина на иностранные языки, написание заимствованного русского термина на иностранных языках, разноязычные эквиваленты античного термина и т.д. Примеры:
I. 1. гекзаметр,
3.1. др.-греч. ἑξάμετρον, 3.2. англ. hexameter, 3.3. нем. Hexameter, 3.4. франц. hexamètre, 3.5. итал. esametro;
II. 1. верлибр,
3. франц. vers libre;
III. 1. дольник,
3. англ. dolnik.

4. «Синонимы». Сюда включаются эквиваленты тезавриуемого термина в русском языке. Синонимия может быть полной и неполной (дискуссионным вопросом для ряда терминов будет включение их в это или в поле «Ассоциации» тезавриуемого термина). Для некоторых терминов содержание поля «Синонимы» может исчерпывать дефиницию, оставляя поле 5 «определение» пустым. Примеры:
I. 1. верлибр,
4. свободный стих;
II. 1. дольник,
4. паузник;
III. 1. хорей,
4. трохей;
IV. 1. эндекасиллаб,
4. одиннадцатисложник.

5. «Теоретическое (дескриптивное) определение» Дефиниция, задающая объект перечислением необходимых его свойств или функций, описание термина на русском языке. Опыт показывает, что дефиницию приходится почти всегда компилировать заново: заимствовать ее в одном из источников тезауруса обычно не удается в силу разных принципов определения термина в тезаурусе и его источниках. Расхождения между предлагаемой дефиницией и иными, существующими в научной традиции, учитываются в поле 6. Примеры:
I. 1. стих$_1$,
5. Художественная речь, фонически расчлененная на относительно короткие отрезки, которые воспринимаются как сопоставимые и соизмеримые;
II. 1. ямб,
5. Стих, состоящий из двусложных стоп с сильным местом на втором слоге и слабым местом на первом;
III. 1. силлаботоника,
5. Система стихосложения, в которой наименьшей стихообразующей парадигматической единицей является стопа.

6. «Альтернативные определения». Для одного термина может быть указано несколько дефиниций. Существует практика вынесения их за пределы ТСТ (в комментарий и т. д.). Представляется, однако, что для тезауруса как средства информационного поиска такие определения должны включаться в основной состав ТСТ. Примеры:
I. 1. стих$_1$,
6. Система сквозных принудительных парадигматических членений.

7. «Конструктивное определение» Явное описание строения соответствующего объекта. Примеры:
I. 1. стих$_1$,
7. отсутствует;
II. 1. ямб,
7. Стих, в котором ударения неодносложных слов могут падать только на четные слоги;
III. 1. силлаботоника,
7. Система стихосложения, основанная на правильном чередовании сильных слогов, по которым распределяются ударения, и слабых слогов, на которые ударения не падают.

8. «Схема (формула)» Для некоторых явлений конструктивное определение может быть представлено в виде схемы или формулы. Примеры:
I. 1. ямб,
8. $\cup$ — $\cup$ — .... $\cup$ $\overset{\prime}{—}$ ($\cap$)($\cap$);
II. 1. хорей,
8. — $\cup$ — $\cup$ .... $\overset{\prime}{—}$ ($\cap$)($\cap$);
III. 1. дактиль,
8. — $\cup$ $\cup$ — $\cup$ $\cup$ .... $\overset{\prime}{—}$ ($\cap$)($\cap$);
IV. 1. дольник,
8. ($\cap$)($\cap$) — $\cup$ ($\cup$) — $\cup$ ($\cup$) .... $\overset{\prime}{—}$ ($\cap$)($\cap$);

Нетрудно видеть, что позиционирование таких полей как «Конструктивное определение» и «Схема (формула)» имеет решающее значение для автоматизированной спецификации поэтических текстов, однако заполнение этих полей представляет отдельную и непростую задачу для специалистов.

9. «Рубрика». В названной выше рубрикации (классификации) – конкретная рубрика, в которую помещен данный темин. Однако определить такую рубрику при заполнении отдельной ТСТ, если термин иерархически не примыкает к уже определенной рубрике верхнего уровня, не всегда возможно. Примеры:
I. 1. метрика,
9. 1.1.1. Метрика;
II. 1. ритмика,
9. 1.1.3. Ритмика.

III. 1. равносложный стих,
4. силлабический) стих,
9. 1.1.1.1 Равносложный (силлабический) стих.

4. Поля ТСТ, содержащие отсылки и комментарии к термину

10. «Дисциплина (рубрика первого уровня)». В описанной выше классификации (рубрикации) рубрика первого уровня. Примеры:
I. 1. ямб,
10. стиховедение;
II. 1. словораздельная вариация,
10. стиховедение;
III. 1. стихомифия,
10. поэтика;
IV. 1. метафора,
10. риторика;
V. 1. частушка,
10. поэтика;
VI. 1. частушка,
10. история литературы.

11. «Национальная традиция». Уточнение отношение термина к национальным терминологическим традициям. Имеется в виду традиция, из которой происходит термин и в которой он первоначально употреблялся, а не традиции, в которых существуют явления, описываемые этим термином. Примеры:
I. 1. гекзаметр,
2. гексаметр,
11. древнегреческая;
II. 1. сатира,
11. древнеримская;
III. 1. пэон,
2.1. пеон, 2.2. пеан,
11. древнегреческая;
IV. 1. эндекасиллаб,
11. итальянская;
V. 1. верлибр,
11. французская.

12. «Автор термина». В тех случаях, когда известен автор, изобретший термин или впервые употребивший нетерминологическое слово в данном терминологическом значении. Примеры:
I. 1. дольник,
12. Брюсов, Валерий Яковлевич (1873–1924)
II. 1. паузник,
12. Бобров, Сергей Павлович (1889–1971)
III. 1. тактовик,
12. Квятковский, Александр Павлович (1888–1968)

13. «Этимология». Происхождение слова (слово языка источника, сведения о его значении и внутренней форме, иногда о его дериватах). Примеры:
I. 1. гекзаметр,
13. шестимерный (др.-греч. ἕξ ‘шесть’ + μέτρον ‘мера’);
II. 1. сатира,
13. лат. satira < satura, от lanx satura ‘полное блюдо’, под влиянием др.-греч. σάτυρος ‘сатир’;
III. 1. пентаметр,
13. пятимерный (др.-греч. πέντε ‘пять’ + μέτρον ‘мера’);
IV. 1. эндекасиллаб
13. одиннадцатисложный (др.-греч. ἕνδεκα ‘одиннадцать’ + др.-греч. συλλαβή, итал. sillabo ‘слог’);
V. 1. дольник,
13. от рус. «доля».
VI. 1. верлибр,
13. свободный стих (франц. vers ‘стих’ + libre ‘свободный’)

14. «Аннотации (статьи)». Резонно помещать в это поле гиперссылки на полные тексты статей, послужившие основными источниками ТСТ. Для печатной версии статьи могут быть использованы эксцерпты (извлечения) из этих источников.

15. «Примеры употребления (цитаты)». Представление исследовательского и контекстного употребления термина: 15.1; 15.2; 15.3.

16. «Источники информации». В качестве источника определения и содержания других полей тезауруса выступают справочно-энциклопедические и словарные издания из предварительно отобранного списка. В основные источники информации могут попадать не все издания из этого списка. (16.1a. библиографическая ссылка; 16.1b. гиперссылка; 16.2a. библиографическая ссылка; 16.2b. гиперссылка).

17. «Дополнительные источники информации». Все источники, использованные для составления ТСТ и не перечисленные в поле 16.

18. «Авторы статьи». Все авторы, участвовавшие в заполнении ТСТ с указанием заполненных ими полей.

Информация этого раздела является важной для последующей рубрикации тезауруса (поле 10) и может быть полезной для построения конструктивного определения термина (поля 13–17).

5. Поля ТСТ, содержащие иерархические отношения термина с другими терминами тезауруса

19. «Родовое понятие». Экспликация (установление) родо-видовых отношений термина (отношение, обратное отношению, фиксируемому в поле 20). У термина может быть только одно родовое понятие. Если их оказывается несколько, мы имеем дело с омонимами, и статья должна быть разбита на две. Примеры:
I. 1. баллада,
19. жанр литературный;
II. 1. амфимакр,
19. стопа;
III. 1. дактиль,
19. стопа трехсложная;

IV. 1. анапест,
19. стопа трехсложная;
V. 1. адоний,
19. логаэд.

20. «Видовые понятия». Экспликация родо-видовых отношений термина (отношение, обратное отношению, фиксируемому в поле 19). У термина может быть несколько видовых понятий (в ряде случаев список видовых понятий является открытым). Примеры:
I. 1. жанр литературный,
20. баллада;
II. 1. стопа,
20. амфимакр;
III. 1. стопа трехсложная,
20. дактиль;
IV. 1. стопа трехсложная,
20. анапест;
V. 1. логаэд,
20. адоний.

21. «Целое». Целое, к которому относится «часть», определяемая тезаврируемым термином (отношение, обратное отношению, фиксируемому в поле 22. Примеры:
I. 1. сапфический одиннадцатисложник,
21. сапфическая строфа;
II. 1. адоний,
21. сапфическая строфа.

22. «Части». Части «целого», определяемого тезаврируемым термином (отношение, обратное отношению, фиксируемому в поле 21. Примеры:
I. 1. сапфическая строфа,
22. сапфический одиннадцатисложник;
II. 1. сапфическая строфа,
22. адоний.

Видимая избыточность приведенных иерархических отношений связана с тем, что при использовании Wiki-технологий для создания тезауруса заполнения полей конкретной ТСТ осуществляется изолированно. Ввиду этого возникает задача последующей автоматизированной рубрикации тезауруса. Наличие в ТСТ связей термина с вышестоящими и с нижестоящими понятиями облегчает построение иерархических ветвей.

Сопоставление содержания иерархических полей ТСТ позволяет определять в первом приближении положение термина в иерархии тезауруса.

1) Если у термина нет ни рода и ни целого и нет ни вида и ни части, то это означает, что термин вне иерархической рубрикации.

2) Если у термина есть род или целое и есть виды или части, то это означает, что термин является промежуточным в иерархической рубрикации.

3) Если у термина есть род или целое и нет ни вида и ни части, то это означает, что термин является конечным в иерархической рубрикации.

4) Если у термина нет ни рода и ни целого и есть виды или части, то это означает, что термин является начальным в иерархической рубрикации.

Эти положения являются основой алгоритма автоматизированной рубрикации, с их помощью можно «сшивать» смежные в рубрикации понятия в отрезки иерархических ветвей, наращивая их до верхнего уровня, после чего можно присвоить рубрики всем узлам построенной ветви. В месте с тем собирание и хранение таких отрезков представляет отдельную информационно-технологическую задачу.

23. При наличии независимых параллельных классификаций родового понятия, приводящих к множественности эквивалентных вариантов построения иерархического дерева. Поэтому для устранения параллелизма в такой комбинативной системе [14] полезно ввести условную иерархию.

23.1. «Ниже». Примеры:
I. 1. равносложный стих,
4. силлабический стих,
19. метрика,
20. отсутствуют,
23.1. равносложный стопный стих.
II. 1. стопный стих,
4. силлаботонический стих,
19. метрика,
20. отсутствуют,
23.1. равносложный стопный стих.

23.2. «Выше». Примеры:
I. 1. равносложный стопный стих,
4. силлаботонический) стих,
19. метрика,
23.2.1. равносложный стих,
23.2.2. стопный стих.

6. Поля ТСТ, содержащие неиерархические отношения термина с другими терминами тезауруса

24. «Соподчинение». Отношения между видами одного рода, в том числе антонимия терминов (23.2 «противоположность» и 23.3 противоречия). Примеры:
I. 1. дактиль,
19. трехсложный размер,
24. амфибрахий, анапест;
II. 1. стих$_1$ ‘форма художественной речи’,
24. проза;
III. 1. проза,
24. стих$_1$ ‘форма художественной речи’.

25. «Смежность». Фиксация метонимических (переносных) отношений термина. Из них исключаются синекдохи («часть-целое» и «целое-часть»), описываемые в полях 21 и 22. Примеры:
I. 1. стих$_1$,
5. художественная речь, фонически расчлененная на относительно короткие отрезки, которые воспринимаются как сопоставимые и соизмеримые,

25. стих$_2$ ‘стихотворная строка’;
II. 1. стих$_2$,
5. стихотворная строка,
25. стих$_1$ ‘форма художественной речи’.

26. «Ассоциации». Все прочие термины, связанные с тезаврируемым термином, отношения с которыми не определяются в полях 23, 24, 25. Такие отношения нередко являются метафорическими. Примеры:
I. 1. стих$_1$ ‘форма художественной речи’,
26. стихотворение в прозе;
II. 1. стихотворение в прозе,
26. стих$_1$ ‘форма художественной речи’.

27. «Комбинативность». Отношение между членами независимых параллельных классификаций родового понятия. Примеры:
I. 1. равносложный стих,
4. силлабический стих,
19. метрика,
20. отсутствуют,
27. стопный стих;
II. 1. стопный стих,
4. силлаботонический стих,
19. метрика,
20. отсутствуют,
27. равносложный стих.

Если какой-то из членов неиерархического отношения имеет конструктивное определение, то это может способствовать конструктивному определению других членов этого отношения.

Неиерархические отношения могут быть приписаны к термину в сводном указателе.

7. Заключение

Представленная логико-семантическая модель «Тезауруса по поэтологии» призвана отвечать требованиям аналитической обработки поэтических текстов и быть системообразующим ядром проектируемой информационно-аналитической системы русской поэзии.

Когнитивная проработка структуры «Тезауруса по поэтологии», его логико-семантической модели и полей ТСТ создают предпосылки для коллективного построения, редактирования и автоматической рубрикации тезауруса с использованием Wiki-технологий.

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.

[2] ГОСТ 7.74-96 СИБИД. Информационно-поисковые языки. Термины и определения. Межгосударственный стандарт. Межгосударственный совет по стандартизации, метрологии и сертификации. Минск

[3] ГОСТ 7.25–2001. СИБИД. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления.

[4] ГОСТ 7.24-2007. СИБИД. Тезаурус информационно-поисковый многоязычный. Состав, структура и основные требования к построению.

[5] Захаров В.Е., Бойков В.Н., Вигурский К.В., Пильщиков И.А. Русская поэзия: проблемы консолидации и анализа в электронном формате. М.: Пробел-2000, 2004.

[6] Лукашевич Н.В. Тезаурусы в задачах информационного поиска. М.: МГУ, 2011.

[7] Нгуен М.Х., Аджиев А.С. Описание и использование тезаурусов в информационных системах, подходы и реализация // Электронные библиотеки. 2004. Вып. 1. URL: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2004/part1/NA .

[8] Никитина С.Е. Тезаурус по теоретической и прикладной лингвистике. М.: Наука, 1978.

[9] Никитина С.Е. Семантический анализ языка науки. На материале лингвистики. М., 1987.

[10] Никитина С.Е., Васильева Н.В. Экспериментальный системный толковый словарь стилистических терминов: Принципы составления и избранные словарные статьи. М.: Институт языкознания РАН, 1999.

[11] Никитина С.Е., Васильева Н.В. Термины лингвистической поэтики в словаре тезаурусного типа // Славянский стих: Стиховедение, лингвистика и поэтика. М.: Наука, 1996. С. 50—57.

[12] Орлов А.Н. Тезаурус информационно-поисковый по литературе, литературоведению, фольклору и фольклористике. М.: ИНИОН АН СССР, 1975.

[13] Русова Н.Ю. От аллегории до ямба: Терминологический словарь-тезаурус по литературоведению. М.: Флинта; Наука, 2004.

[14] Шрейдер Ю.А. Логика классификации. Научно-техническая информация. Сер. 2. 1973. № 5.

[15] ISO 2788:1986. Guidelines for the establishment and development of monolingual thesauri.

[16] ISO 5964:1985. Guidelines for the establishment and development of multilingual thesauri.

A Semantic Model of the “Thesaurus of Poetics” Forming Part of an Information-Analytical System

V.N. Boykov, I.A. Pilshchikov

This paper describes the logical-semantic model of the “Thesaurus of poetics”, which is to become the core of an information-analytical system of Russian poetry (under construction). The model corresponds to the fact that the thesaurus is created as an open network share, and is intended to facilitate the automatic categorization of thesaurus and meet the requirements of the analytical processing of poetic texts.

♣ Работа выполнена при поддержке Российского фонда фундаментальных исследований, грант № 13-06-00448.