ARTSSEMNET: ДВУЯЗЫЧНАЯ СЕМАНТИЧЕСКАЯ СЕТЬ ДЛЯ РУССКОЙ И БОЛГАРСКОЙ ТЕРМИНОЛОГИЙ ИЗОБРАЗИТЕЛЬНОГО ИСКУССТВА

Иванка Я. Атанасова*, Светлин И. Наков, Преслав И. Наков*****

** Великотырновский университет имени Святых Кирилла и Мефодия, Велико-Тырново, Р. Болгария*
*** Софийский университет имени Святого Климента Охридского, София, Р. Болгария*
**** Калифорнийский университет, Беркли, США*

ARTSSEMNET: A BILINGUAL SEMANTIC NETWORK FOR BULGARIAN AND RUSSIAN FINE ARTS TERMINOLOGY

Ivanka Y. Atanassova, Svetlin I. Nakov, Preslav I. Nakov

Abstract: *An electronic lexical reference system ArtsSemNet, similar to WordNet, for terminology of fine arts is presented. The terms (over 2,600 for each language) are annotated with complete dictionary definitions and organized into a semantic network with two parallel versions: Bulgarian and Russian. Five important lexical relations are defined: polysemy, synonymy, homonymy, antonymy and hyponymy, the latter serving as the basis of the hierarchical organization of the ontology. In addition, a specialized browser is created thus providing an intuitive interface to query and navigate through the network.*

Keywords: semantic network, ontology, terminology, polysemy, hyponymy, homonymy, antonymy, synonymy**.**

1. Введение

Повсеместное наступление вычислительных машин оказало большое влияние на современное развитие словарей. Более чем десятилетие назад большинство компьютерно грамотных людей забыло о досадных поисках в больших бумажных словарях и пользуется их компьютерными эквивалентами. Несмотря на то, что первые компьютерные словари во многом уступали классическим, в их потенциальных возможностях никто не сомневался. Еще в 1992 году составители словаря *Oxford English Dictionary* [11] решились инвестировать $13,5 миллионов долларов, чтобы в течение 5 лет построить электронную версию. В то время выяснилось, что электронный вариант словаря предлагает на много больше возможностей. Появились еще тезаурусы (напр. *Roget's thesaurus* [12]), которые предоставляют информацию о синонимах данного термина. Потом лексикографы стали их комбинировать, в результате чего появились *семантические сети* (напр. *WordNet*), которые в терминологии искусственного интеллекта иногда называют *онтологиями*. Они уже включали не только толкования слов и их синонимы, но и антонимы, гипонимы и др.

Таким образом работали и мы – начали с электронных словарей, а потом связали их в полную семантическую сеть посредством терминологических отношений.

2. Семантические сети

WordNet. *WordNet* (в переводе '*сеть слов'*) разработан психолингвистами из Лаборатории когнитивной науки в университете Принстон, США как вычислительная модель человеческой лексической памяти. С течением времени проект эволюировал, превращаясь в лексическую справочную систему с тысячами слов с соответствующими значениями, организованными в семантическую сеть. Словоформы (лексемы) в *WordNet* объединяются во множества, называемые синсетами (от англ. *synset*, что является сокращением от '*синонимического множества'*). Синсет представляет собой объединение слова, обозначающего одно понятие, со значениями других слов (синонимов), чьи лексические значения вместе формируют лексическое значение самого слова [6;9]**.** Многозначные слова участвуют в нескольких различных синсетах, причем каждая отдельная семема включается только в один синсет. Синсеты связаны между собой иерархически согласно реляции гипонимии (с проистекающим оттуда унаследованием) и реляции меронимии, а дальше разграничиваются по различным качествам и свойствам. В *WordNet* (вариант 1.7.1) уже включили 111 223 синсетов – 75 804 имен существительных, 13 214 глаголов, 18 576 имен прилагательных и 3 629 наречий. Проект активен и работа над ним продолжается [14].

EuroWordNet. Вскоре после своего появления *WordNet* вырос как один из важнейших ресурсов для обработки естественного языка, машинного перевода, автоматического определения конкретного значения полисемантичного термина, извлечения информации из текста, извлечения документов в ответ на запрос потребителя и др. В то время как американский *WordNet* развивался, в Европе началась работа над *EuroWordNet* для 7 европейских языков, а именно [13]: голландский, итальянский, испанский, немецкий, французский, чешский и эстонский. Каждая часть *EuroWordNet* построена на основе специфических для конкретного языка синсетов, а все вместе связаны между собой общим индексом на основе *WordNet*, так что возможно переходить между близкими по значению словами различных языков во всех направлениях. Хотя проект *EuroWordNet* [5] был окончен в 1999 (в отличие от *WordNet*, который непрерывно развивается), продолжается работа над различными европейскими языками, а именно: шведский, норвежский, датский, греческий, португальский, баскский, каталонский, румынский, литовский, русский, болгарский и словенский. Позже была создана *Глобальная ассоциация WordNet*, чтобы помогать ученым в дальнейших усилиях в том направлении не только для европейских языков, но и для других современных языков.

MikroKosmos. Конечно, *WordNet* и его иноязычные варианты не являются единственными существенными разработками в этой области. Исторически интересна онтология *MikroKosmos* [7;8], которая была разработана для машинного перевода, однако в настоящее время не используется. Она содержит всего 5 000 терминов, но очень богата отношениями – около 30, включая *IS-A* (гипонимию), *PART-OF* (меронимию), *INSTRUMENT-OF* (инструмент), *LOC-OF* (местоположение) и др.

CYC. Однако, далеко не все онтологии богаты лексическими отношениями. В искусственном интеллекте, например, важнее всего знание о мире (факты), а чтобы описать его вполне хватит одной гипонимии. Так, например *CYC* [4], самая большая

онтология, содержащая около 300 000 терминов и около 3 миллионов фактов о них, создание которой заняло 600 человеко-лет, организована на основе только двух отношений: #$*genls* (подмножество-множество) и #$*is-a* (гипонимия) [4;10].

3. Лексические данные

В основе семантической сети лежит составленный нами софтверный продукт для построения и поддержки компьютерных словарей – *Компьютерный словарь терминов изобразительного искусства* (*КСТИИ*), включающий соответственно 2 644 русских и 2 900 болгарских лексических единиц (вместе с их толкованиями): однословных терминов и терминологических словосочетаний. [1].

Мы исследовали и полностью аннотировали (вручную, но при помощи компьютерных техник) [1; 2] несколько важных терминологических отношений – полисемию, омонимию, синонимию, антонимию и гипонимию. В результате получилась семантическая сеть типа *WordNet,* иерархическая организация которой построена на гипонимии. К моменту изготовления этой статьи семантическая сеть содержит:

- Лексемы в словаре:
  - русские: 2 644 (в том числе абсолютные синонимы, дублеты и варианты);
  - болгарские: 2 900 (в том числе абсолютные синонимы, дублеты и варианты).
- Гипонимические ряды:
  - русские: 226 *(с гиперонимом, включенным в словарь)* + 57 *(с гиперонимом, отсутствующим в словаре)*;
  - болгарские: 216 *(с гиперонимом, включенным в словарь)* + 60 *(с гиперонимом, отсутствующим в словаре).*
- Антонимические ряды:
  - русские: 134;
  - болгарские: 157.
- Синонимические ряды абсолютных синонимов:
  - русские: 458;
  - болгарские: 483.
- Синонимические ряды относительных синонимов:
  - русские: 114;
  - болгарские: 136.
- Омонимы:
  - русские: 6
  - болгарские: 14.
- Полисемия: см. Таблицу № 1.

Количество значений	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Русские термины	2313	263	56	9	2	0	1
Болгарские термины	2571	273	49	4	2	1	0

Таблица № 1. Полисемия терминов в семантической сети.

4. Функциональное описание системы *ArtsSemNet*

Основная задача *ArtsSemNet* – помочь ученому-исследователю в его работе, предоставляя ему способ быстрого и легкого доступа к богатой лингвистической инфор-

мации о терминах изобразительного искусства. При введении конкретного термина *ArtsSemNet* дает информацию об его значениях (толкованиях), омонимах, синонимах (абсолютных и относительных) и синонимических рядах, антонимах и антонимиических рядах, а также о гипонимических рядах, в которые входит термин (в качестве гипонима или гиперонима).

ArtsSemNet предлагает чистый и интуитивный потребительский интерфейс. Потребитель имеет возможность вводить термин в специальное текстовое поле, выбирать язык (болгарский или русский), а также задавать различные критерии для поиска. Система визуализирует наличную информацию о заданном термине на соответствующем языке, включающую:

- различные значения (толкования) термина, последовательно извлекаемые с красной строки;
- список омонимов;
- синонимические ряды *абсолютных* синонимов, состоящие из разнокоренных и однокоренных терминов;
- синонимические ряды *относительных* синонимов;
- антонимические ряды, в которые входит термин;
- гипонимические ряды, возглавляемые введенным термином-гиперонимом;
- гипонимические ряды, в которые входит введенный термин в качестве гипонима.

Система предлагает несколько настроек. Потребитель может задавать дополнительные условия работы: термин можно обнаруживать в *основной* форме подачи или в другой *подобной* форме (например, по корню или префиксу); показывать или пропускать омонимы, синонимы и синонимические ряды, антонимы и антонимические ряды, гипонимы и гипонимические ряды.

Толкования представляют собой текст, объясняющий значения термина на выбранном языке. Значения полисемантичных терминов имеют номера и отделяются друг от друга отступом с красной строки.

Омонимы выводятся в виде списка, в котором каждый термин начинается с красной строки.

Абсолютные синонимы образуют синонимические ряды, в которых термины разделяются между собой тире.

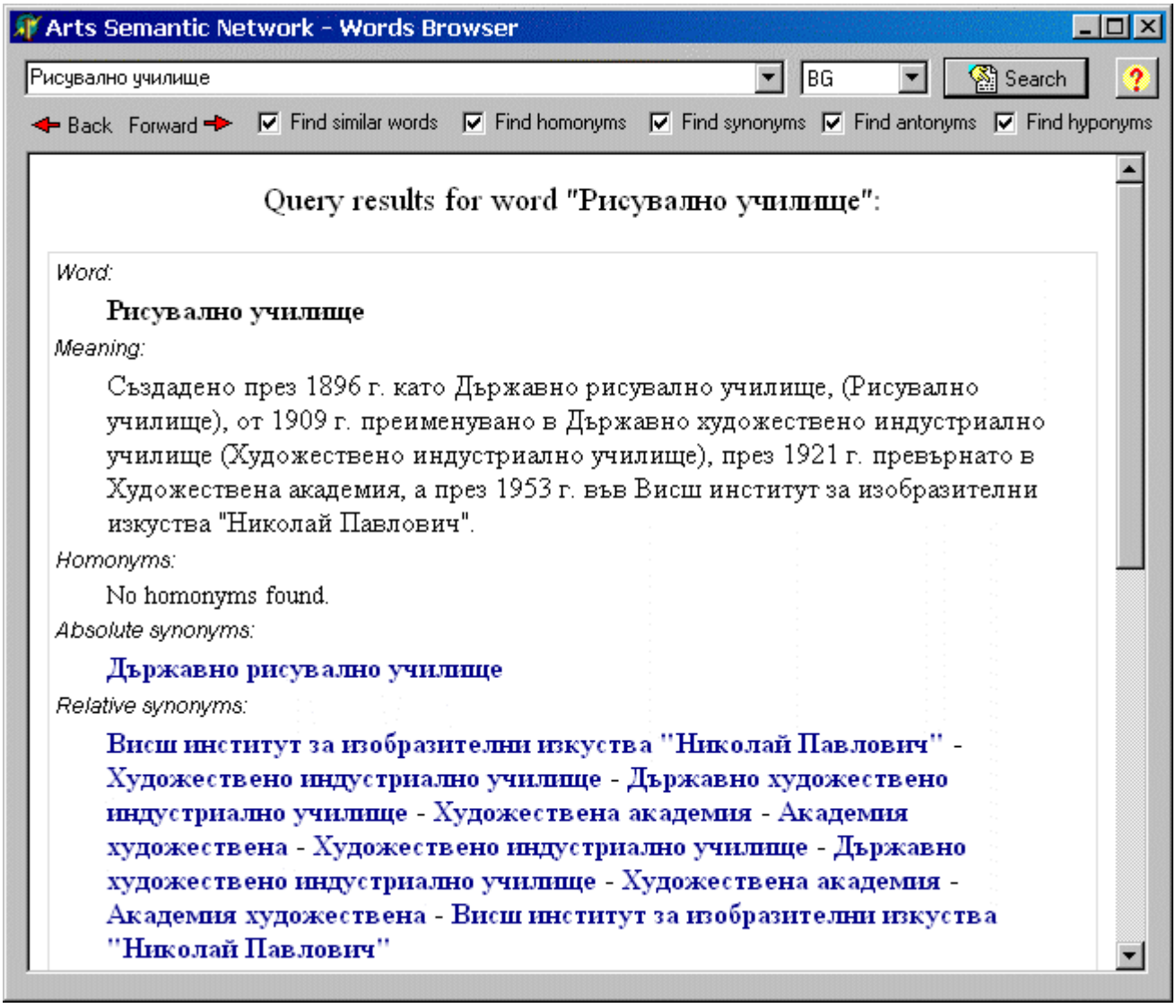
Относительные синонимы тоже образуют синонимические ряды, представляющие собой списки терминов, разделенных тире. Если термин имеет кроме относительных синонимов и абсолютные синонимы, последние стоят рядом с ним, образуя отдельный ряд, члены которого разделяются запятыми.

Антонимы извлекаются в виде антонимических рядов, в которых термины разделяются между собой тире. Рядом с термином-антонимом стоят разделенные запятыми его абсолютные синонимы, которые образуют синонимический ряд.

Гипонимические ряды извлекаются в виде списков терминов, причем первый из них гипероним, возглавляющий гипонимический ряд, а все остальные являются его гипонимами. При наличии абсолютных синонимов у гиперонима или гипонима рядом с ним стоит его синонимический ряд, члены которого разделяются запятыми. Если полисемантичный гипероним возглавляет несколько гипонимических рядов, то после него в скобках указывается соответствующее значение. Этот способ разграничения значений терминов напоминает синсеты в *WordNet*. Потребительский интерфейс дает возможность показывать отдельно каждый гипоним, являющийся гиперонимом для

другого гипонимического ряда, и визуализировать все гипонимические ряды, возглавляемые им.

Во всех случаях, когда извлекаются списки терминов, последние форматируются в виде пересылок (*hyperlinks*), причем осуществляется автоматическая навигация по выбранному термину. После каждой навигации показывается информация о выбранном слове, которая, со своей стороны, тоже может включать пересылки к другим терминам. Выбор термина из пересылок снова визуализирует наличную информацию о нем, и, таким образом, процесс навигации может быть неограниченным. Механизм навигации в терминологии изобразительного искусства, предлагаемый *ArtsSemNet*, очень похож на навигацию в Интернете со стандартным веб браузером. Предусматриваются даже стандартные клавиши для передвижения вперед и назад, изображаемые как левая и правая стрелка. С их помощью потребитель может возвращаться назад в информацию о терминах, которые он рассматривал перед последней навигацией, а потом снова передвигаться вперед при помощи клавиша возврата. Фигура № 1 показывает вид *ArtsSemNet* после удачного поиска болгарского термина *б. рисувално училище*.



Фигура № 1. Вид *ArtsSemNet*.

Система *ArtsSemNet* построена в среде для быстрой разработки приложений *Borland Delphi 6.0*. В целях хранения и извлечения информации о терминах изобразительного искусства использована реляционная база данных *Microsoft Access 2002*, спроектированная таким образом, чтобы обеспечивать быстрое обслуживание всех справок.

5. Доступность *ArtsSemNet*

ArtsSemNet доступна без ограничений для научных исследований, а ее самую актуальную версию (пока только для *Windows*) можно найти в Интернете: http://www.cs.berkeley.edu/~nakov/artssemnet/. По этому адресу можно отыскать и базу данных системы, содержащую всю описанную информацию о терминах изобразительного искусства в русском и болгарском языках, а также и отношения между ними. Распространяется в двух вариантах: 1) *mdb* файл для *Microsoft Access*; и 2) *SQL*-скрипт (создает реляционную схему и заполняет данные в таблицах). Первый вариант удобен для пользования софтверными системами, работающими в *Microsoft Windows*, и вполне доступен даже для потребителей без опыта в работе с реляционными базами данных. Второй вариант может быть использован для пересылок базы данных на *MySQL*, *PostgreSQL*, *Oracle*, *SQLServer* и др. Это дает возможность обрабатывать информацию *ArtsSemNet* с помощью программ, написанных на различных языках программирования, как *Java*, *PHP*, *Perl*, *C#*, *C++* и т. п., включая и работающие по различным операционным системам: *Windows*, *Unix*/*Linux* и др.

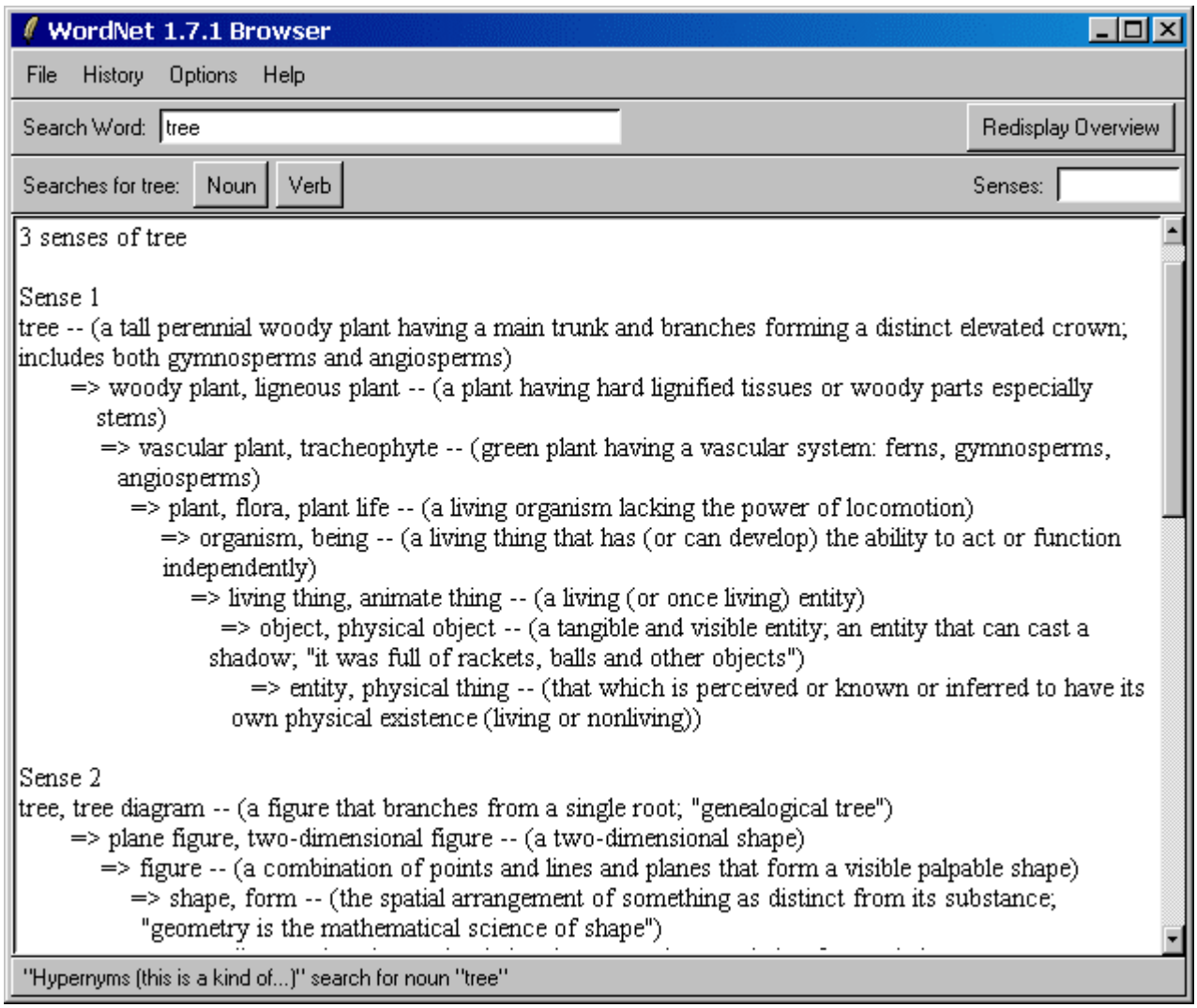
6. *ArtsSemNet* и *WordNet*

WordNet и *ArtsSemNet* имеют сходную функционалность, но между обеими системами существуют и известные расхождения. Как уже было сказано выше, в *WordNet* термины представлены не как самостоятельные слова, а как *синсеты*. *WordNet* построена для английского языка, в котором нередко одно и то же слово может оказаться одновременно именем существительным, именем прилагательным и глаголом. При введении слова для поиска в *WordNet* извлекаются все синсеты вместе с их значениями. Предоставляется возможность для извлечения синонимов, согипонимов, гипонимов и гипонимических рядов, *меронимов* и *голонимов* (термины, в отношении "*X* есть часть *Y*", которые в русском и болгарском языках считаются гипонимами), а также и антонимов в тех случаях, когда слово является именем прилагательным. Все эти сведения касаются соответствующих *синсетов*, связанных с интересующим нас словом, а не *самого слова*. Фигура № 2 показывает результаты поиска гипонимических рядов английского слова *tree* (в переводе *'дерево'*) в системе *WordNet*.

Между *ArtsSemNet* и *WordNet* намечаются следующие основные различия:

- *ArtsSemNet* рассматривает преимущественно самостоятельные термины, а *WordNet* строится на синсетах. *ArtsSemNet* тоже включает подобие синсетов, касающихся, однако, отдельных случаев, главным образом, в связи с представлением гипонимических рядов.
- *ArtsSemNet* поддерживает русский и болгарский языки, а *WordNet* поддерживает только английский язык.
- В отличие от *ArtsSemNet WordNet* не делит синонимы на абсолютные и относительные.

- *WordNet* не рассматривает омонимы, а приводит значения терминов в виде синсетов, причем, если у термина имеется несколько омонимов, то они будут представлены как отдельные синсеты.
- Потребительский интерфейс *WordNet* не предусматривает возможность автоматического наблюдения термина с помощью пересылок, а в *ArtsSemNet* такая возможность предоставляется интуитивным способом, подобным навигации в Интернете.
- *ArtsSemNet* не обнаруживает согипонимы соответствующего термина.
- *ArtsSemNet* не поддерживает меронимию (включает ее в гипонимию).



Фигура № 2. Вид *WordNet*.

7. Будущие разработки

Система *ArtsSemNet* построена специально для справок по русской и болгарской терминологии изобразительного искусства на базе *КСТИИ*, а именно: для извлечения всех значений, омонимов, абсолютных и относительных синонимов и синонимических рядов, антонимов и антонимических рядов, гипонимов и гипонимических рядов заданного термина. Она может быть использована также для сходных конкретных и сопоставительных исследований как терминологических, так и нетерминологических языковых систем.

Существует несколько направлений, в которых система *ArtsSemNet* может развиваться. Во-первых, включение дополнительных справок, например, обнаруживание согипонимов. Другое направление – это построение дополнительных способов для рассмотрения рядов терминов в виде дерева, что создаст более удачное визуальное представление об отношениях между терминами и улучшит навигацию и визуализацию гипонимических рядов. Появится возможность выбора: гипонимические ряды автоматически "развертывать" на месте в главном ряду или показывать в специальном окне при нажиме клавиша. Вполне возможно реализовать отдельную визуализацию в виде дерева для гипонимических рядов, в которой потребитель может осуществлять навигацию.

Другое направление, в котором *ArtsSemNet* может развиваться, это возможность редактирования терминов и связанной с ними информации. Можно будет реализовать функциональность для включения, редактирования и устранения терминов и их значений, омонимов, синонимов и синонимических рядов, антонимов и антонимических рядов, гипонимов и гипонимических рядов. Архитектура системы легко позволяет расширить набор языков, включая кроме русского и болгарского и другие языки. Интересна идея создания межъязыкового индекса наподобие *EuroWordNet*.

Литература

1. Атанасова И. Я., Наков П. И., Наков С. И. Информационные технологии в помощь исследователю-лингвисту. – В кн. Восьмой международный симпозиум МАПРЯЛ 2002. Теоретические и методические проблемы русского языка как иностранного в начале XXI века. Доклады и сообщения. Велико-Тырново, 2002-1, с. 305-307.

2. Атанасова И. Я., Наков П. И., Наков С. И. Семантическая техника автоматического извлечения гипонимических рядов из терминологических словарей. – В кн. Восьмой международный симпозиум МАПРЯЛ 2002. Теоретические и методические проблемы русского языка как иностранного в начале XXI века. Доклады и сообщения. Велико-Тырново, 2002-2, с. 307-313.

3. Новиков, Л. А. Семантика русского языка. М., изд. "Высшая школа", 1982, с. 138;241;142;113;114.

4. CYC, http://www.cyc.com

5. EuroWordNet, http://www.illc.uva.nl/EuroWordNet/

6. Fellbaum C. (ed.). WordNet: An Electronic Lexical Database, MIT Press, 1998.

7. Goodman K., Nirenburg S. (eds.) The KBMT-project: A Case Study in Knowledge-Based Machine Translation. Morgan Kaufmann Publ.,1991.

8. MikroKosmos, http://crl.nmsu.edu/Research/Projects/mikro/index.html

9. Miller G., Beckwith R., Fellbaum C., Gross D., Miller K. Introduction to WordNet: An on-line lexical database. Journal of Lexicography, 3(4), pp. 235-244, 1990.

10. OpenCyc, http://www.opencyc.org

11. Oxford English Dictionary, http://www.oed.com

12. Roget's Thesaurus, http://www.bartleby.com/thesauri

13. Vossen P. (ed.). EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht. 1998.

14. WordNet, http://www.cogsci.princeton.edu/~wn/index.shtml