YARN: НАЧАЛО

Браславский П. И. (pb@kontur.ru)[1, 2]
Мухин М. Ю. (mfly@sky.ru)[1]
Ляшевская О. Н. (olesar@gmail.com)[3, 4]
Бонч-Осмоловская А. А. (abonch@gmail.com)[3, 6]
Крижановский А. А. (andrew.dot.krizhanovsky@gmail.com)[5]
Егоров П. (pe@kontur.ru)[1, 2]

[1]Уральский федеральный университет, Екатеринбург, Россия
[2]Kontur Labs, Екатеринбург, Россия
[3]НИУ Высшая школа экономики, Москва, Россия
[4]ИРЯ РАН им. В. В. Виноградова, Москва, Россия
[5]ИПМИ КарНЦ РАН, Петрозаводск, Россия
[6]МГУ им. М. В. Ломоносова, Москва, Россия

В статье представлен проект создания большого открытого тезауруса русского языка YARN (Yet Another RussNet). Основная особенность проекта — использование wiki-подхода к наполнению и редактированию ресурса. В статье описаны лингвистические принципы создания тезауруса YARN, формат данных, а также ближайшие практические шаги, которые планируется предпринять в рамках проекта.

Ключевые слова: электронный тезаурус, wordnet, лексический ресурс, русский язык

YARN BEGINS

Braslavski P. I. (pbras@yandex.ru)[1, 2]
Mukhin M. Y. (mfly@sky.ru)[1]
Lyashevskaya O. N. (olesar@gmail.com)[3, 4]
Bonch-Osmolovskaya A. A. (abonch@gmail.com)[3]
Krizhanovsky A. A. (andrew.krizhanovsky@gmail.com)[5]
Egorov P. (pe@kontur.ru)[1, 2]

[1]Ural Federal University, Ekaterinburg, Russia
[2]Kontur Labs, Ekaterinburg, Russia
[3]NRU Higher School of Economics, Moscow, Russia
[4]Vinogradov Institute of Russian Language RAS, Moscow, Russia
[5]Institute of Applied Mathematics Research,
Karelian Research Center of RAS, Petrozavodsk, Russia

YARN (Yet Another RussNet) is a work-in-progress on development of a large and open WordNet-like thesaurus for Russian. The paper reports on linguistic design, development and organizational principles, and interchange format of YARN.

Key words: thesaurus, wordnet, lexical resource, Russian language

1. Введение

Большой общедоступный семантический словарь в электронном виде сегодня входит в набор необходимых инструментов и ресурсов для автоматической обработки текстов на конкретном языке (наряду с морфологическим анализатором, синтаксическим парсером, большим аннотированным корпусом и т. п.). Стандартный подход к организации таких ресурсов реализован в проекте Princeton Wordnet (PWN)[1], работа над которым началась в 1986. Сегодня *ворднетами* называют любые лексические базы, построенные по схожим принципам.

PWN состоит из синсетов (*synset,* от *synonym set*) — «смыслов», которые выражаются набором квазисинонимов. В свою очередь синсеты связаны между собой различными отношениями — гипероним-гипоним, мероним-холоним и др. (Рис. 1). В PWN входят существительные, прилагательные, глаголы и наречия. Текущая версия PWN содержит более 117 тыс. синсетов, которым соответствуют примерно 150 тыс. различных словарных входов (отдельных слов и фраз). PWN успешно используется для решения широкого круга задач: снятия лексической неоднозначности, автоматического реферирования,

[1] http://wordnet.princeton.edu/

семантического поиска, классификации и кластеризации документов, обработки поисковых запросов, машинного перевода и т. д.[2]

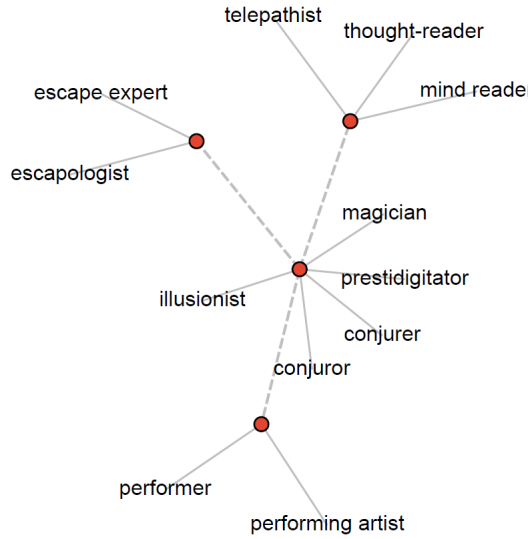

Рис. 1. Наглядное представление фрагмента PWN: слова образуют синсеты (вершины), связи между синсетами (в данном примере — гипо-гиперонимия) обозначены пунктирной линией (http://www.visualthesaurus.com/)

На сегодняшний день wordnet-подобные ресурсы созданы для многих языков, в том числе для малых языков и латыни. Для некоторых языков существует больше одного тезауруса, многие ворднеты являются многоязычными. Таблица на сайте всемирной WordNet-ассоциации содержит сведения о wordnet-подобных ресурсах для 73 языков.[3]

Многие исследователи и практики остро ощущают нехватку общедоступного большого ворднета для русского языка. Поэтому можно сказать, что идея создания открытого большого тезауруса русского языка витала в воздухе. На конференции Диалог-2012 было несколько неформальных обсуждений этой темы, которые продолжились и после конференции. Авторы статьи организовали инициативную группу; проект получил имя YARN (Yet Another RussNet). Кроме ядра группы, представляющего УрФУ (Екатеринбург) и ВШЭ (Москва), к обсуждению проекта подключились участники из Санкт-Петербурга, Казани,

---

[2] Так, запрос [wordnet] на http://aclweb.org/anthology-new/ возвращает ссылки на 4740 статей, а на http://scholar.google.com — на порядок больше (на 10.02.2013).

[3] http://www.globalwordnet.org/gwa/wordnet_table.html

Челябинска, Томска и Петрозаводска. На момент написания статьи у списка рассылки проекта[4] 30 подписчиков.

Основная идея проекта — эксперимент по комбинированию традиционных принципов создания ворднетов и wiki-подхода к наполнению и редактированию лингвистических ресурсов. Данная статья подводит итог примерно полугодовым обсуждениям и подготовительным работам, формулирует основные принципы проекта и — главное — приглашает всех заинтересованных к дискуссии и активному участию в разработке тезауруса.

2. Состояние дел в проблемной области

Принципы разработки PWN (и их историческое развитие), методы автоматического пополнения тезауруса, а также некоторые приложения на основе ресурса описаны в книге [Fellbaum, 1998]. Проект Euro Wordnet[5] [Vossen, 1998] [Alonge et al., 1998] развивал идеи PWN и объединял ворднеты семи европейских языков: голландского, итальянского, испанского, немецкого, французского, чешского и эстонского. Кроме PWN и Euro WordNet на этапе концептуализации YARN мы изучали материалы, связанные с немецким ворднетом GermaNet[6] и польским plWordNet[7]. Обзор других ворднетов можно найти в [Лукашевич, 2011], а также материалах конференций Global WordNet[8] и сайта Global WordNet Association (см. ссылку выше).

Подробнее остановимся на проектах по созданию ворднетов на материале русского языка. Как и в других случаях, мы можем наблюдать два подхода к созданию wordnet-подобных ресурсов для русского языка: 1) перевод и адаптация PWN и 2) создание оригинальной лексической базы данных. Первый подход обеспечивает относительную быстроту заполнения лексикографических баз и более простую интеграцию с аналогичными многоязычными системами. Однако межъязыковые различия, асимметричные явления в лексике препятствуют простому наложению семантической структуры одного языка на лексику другого. Второй подход является более плодотворным с лингвистической точки зрения, но при этом значительно более трудоемким.

С 1999 г. на кафедре математической лингвистики СПбГУ исследовательская группа под руководством И. В. Азаровой ведет работы по проекту RussNet[9] [Азарова и др. 2003]. RussNet ориентирован на описание русской лексики и фиксацию семантических, семантико-грамматических и семантико-деривационных отношений. В тезаурусе отражаются основные типы семантических

---

[4] http://groups.google.com/group/yarn_org/

[5] http://www.illc.uva.nl/EuroWordNet/

[6] http://www.sfs.uni-tuebingen.de/lsd/

[7] http://plwordnet.pwr.wroc.pl/wordnet/

[8] http://www.globalwordnet.org/gwa/gwa_conferences.html

[9] http://project.phil.spbu.ru/RussNet/index_ru.shtml

отношений: синонимия, антонимия, гипонимия, меронимия, конверсия и др. Выявление этих отношений осуществляется с опорой на традиционные лексикографические источники, анализ текстовых корпусов, данные ассоциативных и частотных словарей. По информации на странице проекта, к середине 2005 года тезаурус содержал около 15 тыс. лексико-семантических вариантов, организованных примерно в 5,5 тыс. синсетов, из них 1300 существительных, 1900 глаголов, 1100 прилагательных, 200 наречий. В открытом доступе представлены 300 синсетов «Глаголов эмоционального состояния».

Принципиально другой стратегии построения электронного тезауруса придерживались разработчики Russain WordNet (RWN) из Петербургского государственного университета путей сообщения и ЗАО «Руссикон» [Сухоногов, Яблонский 2005]. Предлагаемая ими версия WordNet является принципиально параллельной англо-русской, т. е. синсеты PWN и отношения между ними переносятся на лексико-семантические варианты русских слов. Для построения и редактирования RWN предлагалось использовать существующие двуязычные словари и автоматические методы. К сожалению, результатов проекта нет в открытом доступе.

Метод автоматического построения русскоязычного тезауруса на основе английского WordNet, корпуса параллельных текстов, двуязычных словарей и словарей синонимов предложила группа исследователей из Новосибирского университета [Гельфенбейн и др. 2003]. По сообщению авторов, с помощью предложенного алгоритма было переведено около 45 % WordNet (авторы пишут о 75 % корректных результатов, однако метод проверки в статье не описан). Большое преимущество проекта состоит в том, что результаты работы в формате принстонского WordNet доступны в сети[10] — тезаурус содержит примерно 18 тыс. существительных, 6 тыс. прилагательных, 5,5 тыс. глаголов, 1,8 тыс. наречий. Однако после знакомства с данными мы пришли к выводу, что проще разрабатывать проект с нуля, чем использовать wordnet.ru как начальный «полуфабрикат» YARN.

Наконец, еще один wordnet-подобный тезаурус русского языка, ориентированный в первую очередь на решение задач информационного поиска, — это РуТез, разработанный в НИВЦ МГУ. По сообщениям авторов, на 2011 РуТез содержал 53 тыс. понятий и 156 тыс. словарных входов. Помимо общеупотребительной лексики, РуТез включает в себя общественно-политические термины. Понятия связаны четырьмя типами отношений: *родо-видовое, ниже — выше*, *часть — целое*, *несимметричная ассоциация* и *симметричная ассоциация*. Тезаурус РуТез, принципы его создания и отличия от PWN подробно описаны в монографии [Лукашевич 2011]. Однако РуТез — это закрытый ресурс, доступный только его создателям.

Альтернативой wordnet-подобным тезаурусам можно считать лексикографические онлайн-ресурсы, в первую очередь — Русский Викисловарь[11]. Кроме

---

[10] http://www.wordnet.ru

[11] http://ru.wiktionary.org/

словарных входов с толкованиями, викисловарь содержит семантические отношения. Очевидные преимущества викисловаря — большой объем и быстрое обновление материала (последнее особенно актуально для неологизмов). Большое количество редакторов, вероятно, обеспечивает более адекватное отражение языковой реальности (противоположностью являются словари, основанные на языковых представлениях небольшой группы лексикографов). Как и другие проекты Web 2.0, викисловарь не может гарантировать качества данных[12], хотя в него и внедряются механизмы, направленные на повышение качества словарных статей[13]. Существуют свободно распространяемые инструменты [Krizhanovsky 2010, Zesch et alt 2008], которые позволяют работать с викисловарем как с базой данных. В последнее время появляются проекты, которые ставят целью сравнить [Meyer&Gurevych, 2012, Смирнов и др. 2012] и интегрировать [Henrich et alt 2011, Navigli&Ponzetto 2012, McCrae et alt 2012, Meyer&Gurevych 2011] ворднеты и викисловари.

Викисловарь по своему замыслу является универсальным ресурсом с максимально расширенным составом словарных зон. На фоне такой масштабной задачи скорость роста и качество его тезаурусной части оставляют желать лучшего. Кроме того, викисловарь не отражает весь набор семантических отношений, выделяемых в wordnet-подобных тезаурусах.

Естественной предпосылкой для создания электронных тезаурусов является большой опыт традиционной идеографической лексикографии. Различные принципы категоризации русской лексики и, соответственно, схемы синопсисов, предложены, например, в Идеографическом словаре О. С. Баранова [Баранов 1995], Русском семантическом словаре под. ред. Н. Ю. Шведовой [Русский семантический словарь, электрон. ресурс], Л. М. Васильева [2003], словарях лексикографической группы под руководством Л. Г. Бабенко, например, в [Большой толковый словарь русских глаголов 2007; Словарь-тезаурус синонимов русской речи 2007], А. Н. Баранова и Д. О. Добровольского [Словарь-тезаурус современной русской идиоматики 2007] и др. Следует заметить, что для разработки структуры wordnet-подобного тезауруса традиционные классификации лексики в чистом виде использовать невозможно. Несмотря на значительные различия между перечисленными источниками, их объединяет общая ориентация на человека-читателя. Ворднеты требуют более высокой степени формализации данных и строгой семантической иерархии с минимальным количеством семантических «вершин». Большое количество уровней классификации не является для электронного тезауруса проблемой, в нем могут порождаться длинные гипо-гиперонимические цепочки синсетов. В результате категоризация лексики осуществляется в основном по направлению от нижних уровней к верхним (от более конкретных по семантике слов — к более абстрактным), а не по пути приписывания слов к готовым семантическим классам.

---

[12] Как замечают [Meyer&Gurevich 2012], «Wiktionary has as yet no reviewing or releasing workflow».

[13] http://ru.wiktionary.org/ Викисловарь:Проверка страниц

3. Лингвосемантические принципы YARN

Разработка большого электронного лингвистического ресурса требует *a priori* продуманных содержательных решений, определяющих представление языковых данных. Следует отметить, что любые форматные ограничения данных связаны с упрощением системных отношений, особенно таких, в которых имеются много нерегулярностей и исключений. Более того, как правило, форматные решения принимаются заранее, и не всегда имеется возможность оценить, насколько часто они будут противоречить интуиции носителя языка. Создание ворднета требует дискретных решений (да/нет) в тех сферах, которые современной теоретической лингвистикой в настоящий момент видятся как континуальные шкалы: что можно считать самостоятельным словом, где границы между одним значением и разными значениями, следует ли конструкции рассматривать как отдельные языковые единицы и т. д.

В основе лингвистических решений, определяющих структуру данных YARN, лежат следующие установки:

1) YARN ориентируется на мировую практику создания wordnet-подобных ресурсов. Первоначальное базовое лексическое наполнение тезауруса будет составлять знаменательная лексика, а именно нарицательные существительные, прилагательные и глаголы. Основу тезауруса составляют синсеты (группы синонимов и квазисинонимов), которые объединены общим лексическим значением. Синсеты упорядочены между собой иерархически и связаны отношениями гиперо/гипонимии (отношение «род — вид» или IS_A в онтологиях) и — далее — антонимии, холонимии/меронимии (отношение «часть — целое» или IS_PART_OF в онтологиях).

Синсеты могут включать в себя словосочетания или отдельные слова; возможен синсет, состоящий из одного слова. Одна и та же лексема может входить в несколько синсетов, каждый такой синсет определяет одно из имеющихся у нее лексических значений.

2) С учетом русской грамматики в тезаурусе приводятся минимальные морфологические характеристики слов, входящих в синсет (имеется ссылка на словоизменительный класс по грамматическому словарю А. А. Зализняка [Зализняк 2010]), проставлено ударение. Омоформы (*печь* сущ. и *печь* гл.) считаются разными лексемами[14]. Разными лексемами также признаются слова, различающиеся своим значением в единственном и множественном числе — например, *выбор* и *выборы*. В этом случае слово типа *выборы* входит в синсет в качестве отдельной лексемы. Словосочетания имеют при себе определение типа связи, отмечены синтаксические вершины. Аббревиатуры всегда состоят в одном синсете с полным наименованием.

3) Учитываются критические замечания к формату PWN (см. об этом [Лукашевич 2011]). В целом, по мере развития ворднеты по своему заполнению

[14] Точнее, они становятся различными входами, если им приписывается грамматическая информация. Такая же ситуация с указанием ударения (замóк и зáмок неразличимы до тех пор, пока не указано ударение).

и формату движутся от словарей к онтологическим тезаурусам. В частности, из них исключаются имена собственные и именованные сущности, применяются более общие принципы выделения синсетов, уточняются параметры наследования признаков (например, меронимических отношений). Наконец, для задачи автоматической обработки языка существенным оказалось отношение domain (предметная область), введенное в третьей версии PWN, по сути своей не лексическое, а онтологическое, т. е. определяющее не значение слова, но объект, обозначаемый этим словом, и относящее его к той или иной области знаний.

Предполагается учесть критику и предложения по совершенствованию PWN, касающиеся отбора лексического материала, гранулярности выделения синсетов и частностей установления семантических отношений между ними.

В перспективе мы планируем интегрировать YARN с разрабатываемым русским Фреймбанком [Ляшевская, Кузнецова 2009, Lyashevskaya 2012]. В самом лексическом ресурсе YARN сохраняются наиболее простые отношения.

4) Отдельную проблему представляет собой организация верхних уровней тезауруса — в силу семантической размытости, многозначности опорных слов и ситуаций словарного взаимоопределения. Очевидный принцип лексикографического конструирования предполагает движение от наиболее конкретных множеств к абстрактным, т. е. построение семантических отношений в направлении снизу вверх (*джинсы* → *брюки* → *одежда* → *товар* → *изделие* → *артефакт* → *вещь*…). Однако такой чисто индуктивный подход к построению тезауруса порождает на верхних уровнях PWN достаточно разнородные семантические парадигмы. См. пример организации существительных (WordNet, ver. 2.1): цепочка {entity}→ {physical entity} → {object, physical object} приводит к 4-му уровню, на котором выделяются 37 единиц: *whole, unit*; *living thing, animate thing*; *location*; *charm, good luck charm*; *curio, curiosity, oddity, oddment, peculiarity, rarity*; *draw, lot*; *film*; *hoodoo*; *je ne sais quoi*; *keepsake, souvenir, token, relic* и т. д. Опыт лексической категоризации подсказывает, что на этом, еще достаточно абстрактном уровне должно выделяться гораздо меньше объектов — например, *вещь, предмет*; *живое существо, существо*; *объект неживой природы*. Поэтому при планировании макроструктуры следует учитывать существующий опыт идеографической лексикографии, т. е. плюсы и минусы уже предложенных синопсисов. Несколько верхних ступеней иерархии будут прописаны в YARN заранее, с тем чтобы конечные пользователи могли с определенностью приписывать к исходящим от вершины объектам гипо-гиперонимические цепочки синсетов.

Макроструктурные перекосы могут быть обусловлены еще и тем, что в wordnet-подобных тезаурусах лексические категории не отделены от слов. Любой публикуемый перечень семантических классов (artifact; attribute; possession; relation etc.) обычно является обобщением выделенных семантических связей. Однако в естественном языке часто отсутствует слово или устойчивый оборот, которые должны соответствовать промежуточному понятию. Например, единица *natural object* (выделяемая в PWN) соответствует в русском языке природному объекту или, еще точнее, объекту неживой природы, которые, строго говоря, устойчивыми оборотами не являются. Наиболее характерна такая ситуация именно для верхних уровней структуры тезауруса. В YARN такие

недостающие смыслы будут восстановлены и станут, наряду с существующими в языке лексическими единицами, полноправными участниками гипо-гиперонимических рядов. Подобные решения заложены в концепции GermaNet и EuroWordNet.

4. Формат

Свободно распространяемый тезаурус используется широким кругом исследователей, встраивается в самые разные приложения, поэтому внешний формат данных такого ресурса — важный элемент проекта. Например, PWN до сих пор использует унаследованный (legacy) формат данных в виде набора текстовых «лексикографических файлов» и скриптов на языке Perl. PWN может «позволить» себе такой подход: на сегодняшний день существует достаточно инструментов и утилит, работающих с этим форматом, или реализующих прозрачный доступ к данным через API. Кроме того, был разработан формат RDF[15] для представления PWN, в результате сообщество semantic web может использовать данные PWN 3.0 в формате RDF[16]. Формат RDF потенциально позволяет легко интегрировать тезаурус в приложения семантического веба. Однако мы согласны с тем, что форматы семантического веба (RDF и построенные на его основе OWL[17] и SKOS[18]) являются «неродными» для лингвистических данных.

Мы считаем, что хорошо структурированные данные в формате XML являются достаточным минимумом, который позволяет, с одной стороны, легко переводить данные в нужный формат и/или использовать отдельные элементы данных, адаптировать под конкретное приложение, а с другой — оставляет связь с внутренним представлением данных в СУБД, не перегружает представление.

Предлагаемый нами XML формат[19] является модульным: разные типы объектов описаны в отдельных частях. Этим предлагаемый формат отчасти напоминает формат Lexical Markup Framework (LMF)[20], хотя и разрабатывался без оглядки на него. Формат учитывает особенности wiki-подхода при редактировании и пополнении тезауруса; информация о редактировании кодируется по аналогии с форматом проекта OpenStreetMap[21].

---

15 http://www.w3.org/TR/wordnet-rdf/

16 http://semanticweb.cs.vu.nl/lod/wn30/

17 http://www.w3.org/2001/sw/wiki/OWL

18 http://www.w3.org/2001/sw/wiki/SKOS

19 https://github.com/xoposhiy/yarn

20 http://www.lexicalmarkupframework.org/

21 http://www.openstreetmap.org/

```
<wordEntry id="n123" approvedBy="mfly" approvedWhen="2012-12-
26T17:14:00Z" author="pb" version="2" timestamp="2012-12-26T17:12:00Z">
        <word>престидижитатор</word>
        <grammar>1a</grammar>
        <accent>12</accent>
        <url>http://ru.wiktionary.org/wiki/престидижитатор</url>
</wordEntry>
```

Рис. 2. Пример описания слова в формате YARN

```
<synsetEntry id="sn1" approvedBy="olesar" approvedWhen="2012-
12-26T17:14:00Z" author="bonch" version="2"
timestamp="2012-12-25T17:12:00Z">
        <word ref="n123">
                <mark>устар</mark>
                <sample url="http://artbrus.livejournal.com/204315.
html">Лучший фокус, когда престидижитатор (так называют фокус-
ника, манипулятора с вещами) начинает вынимать шар, то изо рта,
то из уха, то из-за шиворота обалдевшего зрителя.</sample>
<sample source="В. В. Крестовский, 'Петербургские трущобы', 1867 г.,
НКРЯ">Мечут же карты, передёргивают и всякие иные фокусы употре-
бляют только главные и самые искусные престидижитаторы, которые по-
этому специально называются «дергачами".</sample>
        </word>
        <word ref="n5678"/> <!--манипулятор-->
        <definition url="http://ru.wiktionary.org/wiki/престидижи-
татор" source="wiktionary">фокусник, отличающийся ловкостью рук;
манипулятор</definition>
        <definition source="Толковый словарь Ушакова">фокусник
с большой быстротой и ловкостью рук</definition>
</synsetEntry>
```

Рис. 3. Пример описания синсета в формате YARN

Логически тезаурус состоит из 1) словаря, 2) синсетов и 3) связей.

Первый раздел — это «орфографический словарь», который содержит слова с минимумом дополнительной информации (Рис. 2). Как указывалось выше, в отличие от других тезаурусов, мы предусмотрели (факультативные) поля для грамматических характеристик слов и ударения. Для удобства чтения данных идентификаторы существительных начинаются с *n,* прилагательных и глаголов — с *a* и *v*; аналогично идентификаторы синсетов и связей имеют буквенные префиксы.

Второй раздел тезауруса — синсеты (Рис. 3). У синсета есть только один обязательный элемент — определение (их может быть несколько). Синсет — это набор слов, у слова в контексте синсета появляется пример использования и словарная помета. Синсеты могут быть «пустыми» (не содержать слов), т. е. соответствовать нелексикализованным концепциям, которые вводятся для упорядочения структуры тезауруса (по аналогии с GermaNet).

Наконец, третий раздел — это различные отношения между элементами тезауруса. Основные отношения — это отношения между синсетами (см. выше). Дополнительно есть отношения между словами (по аналогии с crossPOS-отношениями в PWN), а также межъязыковые связи между синсетами YARN и PWN (соответствуют записям межъязыкового индекса в Euro Wordnet, Global Wordnet Grid[22] и Open Multilingual Wordnet[23]). Особый тип отношения — антонимия: отношение между словами в контексте конкретных синсетов.

Каждый элемент данных имеет атрибуты, позволяющие отслеживать авторство и версионность (имена атрибутов говорят сами за себя):

- id;
- approvedBy;
- approvedWhen;
- author;
- version;
- deleted;
- timeStamp.

Например, выборка всех элементов с атрибутом *deleted=false*, а среди объектов с одинаковым *id* выбор того, чья версия максимальна, возвращает нам последнюю версию тезауруса. Если добавить условие «атрибут *approvedBy* не пуст», то получится последняя версия тезауруса, проверенная редакторами.

5. Принципы организации и первые шаги

5.1. Распространение данных проекта

Планируется распространять данные YARN по лицензии Creative Commons «Attribution-ShareAlike» («Атрибуция — На тех же условиях»)[24] (по этой лицензии распространяется контент Wikipedia[25]). Лицензия разрешает использовать материалы в исследовательских и коммерческих приложениях, модифицировать и перераспространять данные и требует только ссылаться на источник. Мы планируем регулярно публиковать текущий дамп YARN и предоставлять свободный и бесплатный доступ к данным.

---

[22] http://www.globalwordnet.org/gwa/gwa_grid.html

[23] http://casta-net.jp/~kuribayashi/multi/

[24] http://creativecommons.org/licenses/by-sa/3.0/

[25] http://wikipedia.org

5.2. Wiki-принцип наполнения и редактирования

В отличие от существующих тезаурусов, мы не хотим ограничивать наполнение и редактирование ресурса рамками одной исследовательской группы. Все данные должны храниться централизованно и редактироваться через веб-интерфейс. В то же время мы предполагаем контроль процесса редактирования и качества. Так, в отличие от википедии и викисловаря, редактирование данных незарегистрированными пользователями не будет разрешено. Среди пользователей будет выделено ядро редакторов, которые могут утверждать или отменять изменения, а также запрещать дальнейшее редактирование отдельных элементов тезауруса. Мы рассчитываем, что сможем привлечь к редактированию YARN большое количество участников[26].

5.3. Инструмент редактирования

Потенциальные участники — люди с разными навыками, предпочтениями и уровнем лингвистических знаний. Поэтому особенно важно предоставить им удобный и простой интерфейс для совместной работы. Мы не нашли свободно распространяемого лексикографического инструмента, который соответствовал бы таким требованиям или мог бы быть доработан, поэтому начали работу по разработке собственного инструмента редактирования тезауруса.

5.4. Викисловарь — источник данных на начальном этапе

На первом этапе планируется использовать данные викисловаря — свободно пополняемого и распространяемого многофункционального источника. Структура словарной статьи викисловаря определяется правилами[27]; с позиций нашего проекта наибольший интерес представляют следующие разделы:

- *Морфологический раздел*. В нём указаны морфологические свойств; членение слова на морфемы.
- *Семантический раздел*. Включает толкования и цитаты, иллюстрирующие каждое из значений слова. Цитаты сопровождаются библиографической информацией: автор, название произведения, год издания, источник (корпус текстов или электронная библиотека). С помощью помет определяется сфера использования слова (литературная, диалектная, терминологическая, жаргонная лексика), предметная область (авиационная, автомобильная, альпинистская и т. д.). Также в этом разделе описываются семантические отношения (синонимы, гиперонимы и т. д.) отдельно для каждого из значений слова.

---

[26] Оценка основана на опыте проекта http://opencorpora.org/

[27] См. http://ru.wiktionary.org/wiki/Викисловарь:Правила_оформления_статей.

- *Раздел родственных слов* группирует однокоренные слова с разбиением по частям речи (отдельно указываются, например, уменьшительно-ласкательные формы).

По данным [Смирнов и др., 2012] в Русском Викисловаре на весну 2011 года было 135 396 входов, из них — 53 635 слов с толкованиями. Распределение слов по интересующим нас частям речи приведено в табл. 1.[28]

Табл. 1. Число русских слов и значений по частям речи

Часть речи	Лексема (уникальная строка)	Кол-во слов с одним значением	Всего значений
Существительное	22 281	14 718	35 320
Глагол	6 654	2 585	16 637
Прилагательное	5 288	3 091	9 069
Наречие	1 215	822	1 805

На март 2012 года для слов русского языка в Русском Викисловаре указано: синонимов — 56 844, антонимов — 17 047, гиперонимов — 34 166, гипонимов — 13 751, холонимов — 538, меронимов — 910. Динамика количества синонимических ссылок русского викисловаря приведена в Табл. 2.

Табл. 2. Скорость роста числа русских словарных статей и ссылок на синонимы в Русском Викисловаре

	2010	2011	2012
Входы	83 968	135 396 (+61 %)	158 689 (+17,2 %)
Синонимические ссылки	36 158	47 009 (+30 %)	56 844 (+20,9 %)

Полноценный лексикографический анализ качества словарных статей Викисловаря на данный момент не осуществлен. Есть ряд работ [Meyer, Gurevich 2012], [Смирнов и др. 2012], где выполнено сравнение нескольких викисловарей и WordNet методами квантитативной лингвистики и получен вывод, что словари, разрабатываемые энтузиастами, демонстрируют те же закономерности, что и традиционные словари, созданные специалистами. Тем не менее ожидается, что в ходе разработки YARN удастся преодолеть основные недостатки викисловарей: (1) отсутствие проверки целостности данных (например, можно ввести несуществующий код языка), 2) амбивалентность синонимических ссылок, направленых на словарные статьи, а не на конкретные значения

[28] Одна словарная статья в Викисловаре может включать несколько омонимов, например, частеречные омонимы: глагол «течь» (протекать) — шесть значений и существительное «течь» (протекание) — два значения. В этом случае вклад словарной статьи «течь» в таблицу будет равен одному глаголу и одному существительному (лексемы), шести значениям для глаголов и двум — для существительных (значения).

в статье. Первый недостаток можно устранить, создав достаточно жесткую инструкцию для краудсорсинга и прозрачную систему редакторского контроля качества. Планируется также использование современных возможностей онлайн-оценки качества данных («wisdom of the crowd»). Устранение второго недостатка заложено в самой идеологии ворднета.

5.5. Первые практические шаги

Ближайший этап наполнения YARN — формирование синсетов существительных из викисловаря. Поток заданий будет сформирован на основе частотного словаря НКРЯ (для того, чтобы отсечь, например, входы http://ru.wiktionary.org/wiki/сепулька или http://ru.wiktionary.org/wiki/брисбенец). На текущем этапе осуществляется разработка основных интерфейсов и создание инструкции по заполнению тезауруса. Интерфейсы и инструкция пройдут тестирование в узком круге редакторов (преимущественно студентов-лингвистов). На следующем этапе тезаурус будет размещен в открытом доступе для привлечения широкого круга пользователей. Как было сказано выше, данные, внесенные с помощью краудсорсинга, будут проходить редакторский контроль.

Саму процедуру заполнения синсетов в общем виде можно представить следующим образом:

На вход пользователю поступает список слов, в котором отображается статус каждого слова: включено ли слово уже в какой-либо синсет.

Для работы над синсетом пользователь опирается на три типа входных данных: а) толкования слова, полученные из разных источников (Викисловарь и ряд толковых словарей); б) синонимы к слову; в) контекстные примеры на употребление слова (НКРЯ[29]).

Анализируя данные, пользователь определяет синсеты, в которые входит слово, формирует их толкования, подбирает и добавляет в синсет синонимы, верифицирует синонимические ряды с помощью реальных текстовых примеров.

Каждый из синсетов пользователь привязывает к гипероним, а также при необходимости формирует гипонимически подчиненный синсет. Исходя из этого, отдельной задачей главных редакторов YARN станет оптимизация гипо-гиперонимических связей.

Таким образом, основная стратегия краудсорсинга YARN состоит, во-первых, в экстенсивном расширении включаемых лексем, а во-вторых, в итеративном редактировании устанавливаемых связей. Организация работы над русским ворднетом подразумевает иерархию редакторских полномочий, с одной стороны, и использование социальных возможностей ресурса открытого доступа для оценки качества синсетов и отношений, с другой.

[29] http://ruscorpora.ru

6. Заключение

Коллектив разработчиков приглашает к сотрудничеству любых экспертов (лингвистов и программистов), — в особенности, специалистов, имеющих опыт разработки тезаурусов на материале русского языка. Организационные и технологические особенности YARN позволяют интегрировать различные лингвистические ресурсы.

YARN — проект, который находится в стадии становления и определения концептуально-лингвистических и чисто технических установок. В то же время группа разработчиков придерживается принципов, которые определены давно назревшей необходимостью создания русского ворднета, а также опытом организации подобных ресурсов и идеографических словарей. Это максимальное расширение числа участников редактирования словарных данных, привлечение возможностей автоматизированного извлечения лексикографической информации и открытость проекта в плане его дальнейшего развития и использования. Мы полагаем, что только такие принципы смогут наконец привести к созданию лингвистически обоснованного свободно распространяемого электронного тезауруса русского языка.

Благодарности

Исследование осуществляется при финансовой поддержке РГНФ (проект № 13-04-12020 «Новый открытый электронный тезаурус русского языка»).

Мы благодарим участников группы *yarn_org* за активность, замечания и предложения.

Работа Андрея Крижановского выполнена при частичной финансовой поддержке РФФИ (проект № 11-01-00251, № 12-01-00481, № 12-07-00070) и РГНФ (проект № 12-04-12062).

Работа Ольги Ляшевской и Анастасии Бонч-Осмоловской отражает результаты исследований, проведенных при поддержке Программы фундаментальных исследований НИУ Высшая школа экономики (2013), проект «Корпусные технологии в лингвистических и междисциплинарных исследованиях».

Павел Браславский благодарит группу разработчиков GermaNet под руководством проф. Эрхарда Хинрихса из университета Тюбингена за гостеприимство, плодотворное обсуждение проекта и обмен опытом, а также MUMIA Network[30] за финансовую поддержку визита в Тюбинген в рамках программы Short Term Scientific Missions (STSM).

[30] http://www.mumia-network.eu

Литература

1. *Азарова И.В., Митрофанова О. А., Синопальникова А. А.* (2003). Компьютерный тезаурус русского языка типа WordNet. Диалог-2003. Протвино, Россия.
2. *Баранов О. С.* Идеографический словарь русского языка. М.: ЭТС, 1995.
3. *Большой толковый словарь русских глаголов:* Идеографическое описание. Синонимы. Антонимы. Английские эквиваленты / под ред. Л. Г. Бабенко. М.: АСТ-Пресс книга, 2007.
4. *Васильев Л. М.* Системный семантический словарь русского языка. Предикатная лексика. Ментальные предикаты. Модальные предикаты. Предикаты восприятия. Уфа: РИО БашГУ, 2003.
5. *Гельфейнбейн И. Г., Гончарук А. В., Лехельт В. П., Липатов А. А., Шило В. В.* (2003). Автоматический перевод семантической сети WordNet на русский язык. Труды Международного семинара Диалог по компьютерной лингвистике и её приложениям, Протвино, Россия.
6. *Зализняк А. А.* Грамматический словарь русского языка. М.: Русский язык, 1977.
7. *Лукашевич Н. В.* Тезаурусы в задачах информационного поиска, М.: Изд-во МГУ, 2011.
8. *Ляшевская О. Н., Кузнецова Ю. Л.* Русский фреймнет: к задаче создания корпусного словаря конструкций // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). М.: РГГУ, 2009. С. 306–312.
9. *Русский семантический словарь* / под общей ред. Н. Ю. Шведовой. http://www.slovari.ru/default.aspx?s=0&p=235
10. *Словарь-тезаурус синонимов русской речи* / под общ. ред. Л. Г. Бабенко. М.: АСТ-Пресс книга, 2007.
11. *Словарь-тезаурус современной русской идиоматики* / под ред. А. Н. Баранова, Д. О. Добровольского. М. : Мир энциклопедий Аванта+, 2007.
12. *Смирнов А. В., Круглов В. М., Крижановский А. А., Луговая Н. Б., Карпов А. А., Кипяткова И. С.* Количественный анализ лексики русского WordNet и викисловарей // Труды СПИИРАН. 2012. Вып. 23. С. 231–253. http://scipeople.com/publication/113406/
13. *Сухоногов А. М., Яблонский С. А.* Автоматизация построения англо-русского Wordnet. Диалог-2005.
14. *Alonge, A., Calzolari, N., Vossen, P., Bloksma, L., Castellon, I., Marti, M., & Peters, W.* (1998), The Linguistic Design of the EuroWordNet Database. Computers and the Humanities, Vol. 32(2–3), 91–115.
15. *Fellbaum, C.* WordNet: An Electronic Lexical Database, Mit Press, Cambridge, 1998.
16. *Henrich V., Hinrichs E., Vodolazova T.* (2011), Semi-Automatic extension of GermaNet with sense definitions from Wiktionary. Proceedings of 5th Language & Technology Conference (LTC 2011). Poznan, Poland, pp. 126–130, available at: http://www.sfs.uni-tuebingen.de/lsd/documents/publications/Henrich-et-al-2011_GermaNet-Wiktionary-Mapping.pdf

17. *Krizhanovsky A. A.* (2010), Transformation of Wiktionary entry structure into tables and relations in a relational database schema? available at: http://arxiv.org/abs/1011.1368
18. *Lyashevskaya O. N.* (2012), Dictionary of Valencies Meets Corpus Annotation: A Case of Russian FrameBank, Proceedings of EURALEX 15, Oslo.
19. *McCrae J., Montiel-Ponsoda E., and Cimiano Ph.* (2012), Integrating WordNet and Wiktionary with lemon, Conference Proceedings "Linked Data in Linguistics", pp. 25–34.
20. *Meyer Ch. M., Gurevych, I.* (2011), What psycholinguists know about chemistry: Aligning Wiktionary and wordnet for increased domain coverage, Proceedings of the 5th international joint conference on natural language processing (IJCNLP), Chiang Mai, Thailand, pp. 883–892.
21. *Meyer Ch. M., Gurevych I.* (2012), Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography, chapter 13, in Sylviane G., Paquot M. (eds.), Electronic Lexicography,Oxford University Press, Oxford, pp. 259–291.
22. *Navigli R, Ponzetto S. P.* (2012), BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, Vol. 193, pp. 217–250.
23. *Vossen, P.* EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Springer, 1998.
24. *Zesch.T., Müller Ch., and Gurevych I.* (2008), Extracting lexical semantic knowledge from Wikipedia and Wiktionary, Proceedings of the Conference on Language Resources and Evaluation (LREC). Vol. 15.

References

1. *Alonge, A., Calzolari, N., Vossen, P., Bloksma, L., Castellon, I., Marti, M., & Peters, W.* (1998), The Linguistic Design of the EuroWordNet Database. Computers and the Humanities, Vol. 32(2–3), 91–115.
2. *Azarova I. V., Mitrofanova O. A., Sinopal'nikova A. A.* (2003), Russian Computational Thesaurus [Komp'juternyj tezaurus russkogo yazyka], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2003" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2003"], Protvino.
3. *Baranov O. S.* (1995) The Ideographic Dictionary of Russian [Ideograficheskij slovar› russkogo jazyka], ETC, Moscow.
4. *The Big Explanatory Dictionary of Russian Verbs* (2007): Ideographical Description. Synonyms. Antonyms. English Equivalents [Bol'shoj ideograficheskij slovar' russkih glagolov: Ideograficheskoje opisanije. Sinonimy. Antonimy. Anglijskije ekvivalenty], edited by L. G. Babenko, AST-Press Kniga, Moscow.
5. *Fellbaum, C.* WordNet: An Electronic Lexical Database, MIT Press, Cambridge, 1998.
6. *Gel'fejn'bejn I. G., Goncharuk A. V., Lehel't V. P., Lipatov A. A., Shilo V. V.* (2003), Automatic translation of Wordnet in russian [Avtomaticheskij perevod semanticheskoj seti Wordnet na russkij jazyk], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2003" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2003"], Protvino.
7. *Henrich V., Hinrichs E., Vodolazova T.* (2011), Semi-Automatic extension of GermaNet with sense definitions from Wiktionary, Proceedings of 5th Language & Technology Conference (LTC 2011), Poznan, Poland, pp. 126–130, available at: http://www.sfs.uni-tuebingen.de/lsd/documents/publications/Henrich-et-al-2011_GermaNet-Wiktionary-Mapping.pdf
8. *Krizhanovsky A. A.* (2010), Transformation of Wiktionary entry structure into tables and relations in a relational database schema, available at: http://arxiv.org/abs/1011.1368
9. *Lukashevich N. V.* (2011), Thesauruses in Information Retrieval [Tezaurusy v zadachax informacionnogo poiska], Izdateljstvo MGU, Moscow
10. *Lyashevskaya O. N.* (2012), Dictionary of Valencies Meets Corpus Annotation: A Case of Russian FrameBank, Proceedings of EURALEX 15, Oslo.
11. *Lyashevskaya O. N., Kuznetsova Ju. N.* (2009), Russian Framenet: towards the development of Russian corpora constructional thesaurus [Russkij Framenet, k zadache sozdanija korpusnogo slovar'a konstruktsyj], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2009" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2009"], Bekasovo.
12. *McCrae J., Montiel-Ponsoda E., and Cimiano Ph.* (2012), Integrating WordNet and Wiktionary with lemon, Conference Proceedings "Linked Data in Linguistics", pp. 25–34.

13. *Meyer Ch. M., Gurevych, I.* (2011), What psycholinguists know about chemistry: Aligning Wiktionary and wordnet for increased domain coverage, Proceedings of the 5th international joint conference on natural language processing (IJCNLP), Chiang Mai, Thailand, pp. 883–892.
14. *Meyer Ch. M., Gurevych I.* (2012), Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography, chapter 13, in Sylviane G., Paquot M. (eds.), Electronic Lexicography,Oxford University Press, Oxford, pp. 259–291.
15. *Navigli R., Ponzetto S. P.* (2012), BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artificial Intelligence, Vol .193, pp 217–250.
16. *Russian Semantic Dictionary* [Russkij Semanticheskij slovar'], edited by N. Yu. Shvedova available at: http://www.slovari.ru/default.aspx?s=0&p=235
17. *Smirnov A. V., Kruglov V. M., Krizhanovskiy A. A., Lugovaja N. B., Karpov A. A., Kip'atkova I. S.* (2012), The quantative analysis of lexics in Russian Wordnet and wikidictionaries [Kolichestvennyj analiz leksiki russkogo Wordnet i vikislovarej], Proceedings of SPIIRAN, Vol. 23, pp. 231–253.
18. *Suhonogov A. M., Jablonskij S. A.* (2005), Automatic construction of English-Russian Wordnet [Avtomatizacija postroenija anglo-russkogo Wordneta], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2005" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2005"], Bekasovo.
19. *The Thesaurus of Russian Synonyms* (2007) [Slovar'-tezaurus sinonimov russkoj rechi], edited by L. G. Babenko, AST-Press Kniga, Moscow.
20. *The Thesaurus of Contemporary Russian Idiomatic Expressions* (2007) [Slovar'-tezaurus sovremennoj russkoj idiomatiki], edited by A. N. Baranov & D. O. Dobrovol'skij, Mir Encyclopedij Avanta+, Moscow.
21. *Vasiljev L. M.* (2003) The System Semantic Dictionary of Russian. Predicates. Mental Predicates. Modal Predicates. Predicates of Perception [Sistemnyj semanticheskij slovar' russkogo jazyka. Predikatnaya Leksika. Mental'nyje predikaty. Modal'nyje predikaty. Predikaty vospriyatiya], BashGU, Ufa.
22. *Vossen, P.* EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Springer, 1998.
23. *Zalizn'ak A. A.* (1977), Grammatical Dictionary of Russian [Grammaticheskij slovar' russkogo jazyka], Russkij Jazyk, Moscow.
24. *Zesch T., Müller Ch., and Gurevych I.* (2008), Extracting lexical semantic knowledge from Wikipedia and Wiktionary, Proceedings of the Conference on Language Resources and Evaluation (LREC). Vol. 15.