Подход к разработке русско-английского тезауруса по компьютерной лингвистике♣

© Ю.А. Загорулько[1], О.И.Боровикова[1], И.С. Кононенко[1], Е.Г. Соколова[2]

[1]Институт систем информатики имени А.П.Ершова СО РАН, г. Новосибирск
[2]Российский государственный гуманитарный университет, г. Москва
zagor@iis.nsk.su, olesya@iis.nsk.su, irina_k@cn.ru, minegot@rambler.ru

Аннотация

В докладе представлен подход к разработке русско-английского электронного тезауруса по компьютерной лингвистике. Рассматривается общее строение тезауруса, структура тезаурусных статей и набор связей между терминами. Обсуждается проблема выбора терминов для включения в тезаурус, а также проблема выбора основного термина-дескриптора из множества синонимичных терминов. Описываются особенности реализации электронной версии тезауруса, при этом особое внимание уделяется поддержанию логической целостности терминологической системы тезауруса и обеспечению удобного доступа к его содержимому.

1. Введение

В настоящее время наблюдается значительный интерес к компьютерной лингвистике (КЛ), как к прикладной научной дисциплине, включающей знания о методах извлечения информации из текстов, индексирования и содержательного поиска документов, построения естественно-языковых, в том числе речевых, интерфейсов. В связи с этим возникла острая потребность в систематизированных знаниях по терминологии КЛ, которые, с одной стороны, способствовали бы повышению образовательного уровня, а с другой стороны, использовались для индексирования публикаций по КЛ – как ручного/автоматизированного, так и автоматического – с целью облегчения доступа к представленным в них знаниях по КЛ.

Однако в данный момент в КЛ отсутствует четкая и общепринятая система научной терминологии, причем многие термины современной КЛ не представлены на русском языке ни в одном из существующих лингвистических источников.

Так, тезаурус по теоретической и прикладной лингвистике, созданный в 1978 г. С.Е. Никитиной [9], уже устарел. Кроме того, он одноязычный и не содержит определений понятий. Терминологический словарь В.З. Демьянкова [6] содержит толкования и является двуязычным, но не отражает современную картину этой научной области.

Представительного компактного собрания терминов современной КЛ и их толкований не существует не только в России, но и за рубежом. Собственно лингвистика представлена в нескольких фундаментальных источниках, в частности, в ЛЭС [10], словаре О.С. Ахмановой [1] и интернет-энциклопедии «Кругосвет» [13], содержащей статьи по новым для традиционной лингвистики понятиям. Разработанный в 2007 г. в ИНИОН РАН тезаурус по языкознанию [19] содержит около 3000 терминов, относящихся к различным разделам данной науки. При всех своих достоинствах, данный тезаурус, прежде всего, предназначен для библиографического поиска, поэтому его словарные статьи не содержат дефиниций. Кроме того, тезаурус ИНИОН является одноязычным и характеризуется малым удельным весом собственно терминологии КЛ (дескрипторы из области КЛ составляют около 4% от общего количества терминов, представленных в тезаурусе).

Определения терминов КЛ содержатся и в толковом словаре по искусственному интеллекту [18]. Однако он отражает терминологию на конец 1980-х гг. Кроме того, он содержит довольно мало терминов КЛ, а имеющиеся в нем термины чаще всего трактуются не с позиций этой области знаний, а с позиций искусственного интеллекта.

Так как КЛ имеет междисциплинарный характер, то некоторые ее термины можно найти в общих энциклопедиях, например, в БЭС [2]. Популярным источником знаний по КЛ сейчас является Википедия [8], в которой можно найти объяснения, классификации и ссылки на источники по многим понятиям КЛ, однако эти сведения часто страдают односторонностью, неполнотой и эскизностью.

Труды 13й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

Таким образом, на данный момент не существует источника, в котором вся терминология КЛ была бы приведена в единую систему. Этот ощутимый и досадный пробел мог бы восполнить двуязычный тезаурус, содержащий английские и русские термины КЛ и их толкования.

Такой тезаурус позволит структурировать и накапливать информацию, релевантную для этой области знаний, производить смысловой поиск данных в информационных хранилищах и сетях. Кроме того, такой тезаурус мог бы способствовать повышению уровня профессиональной подготовки будущих специалистов не только в сфере КЛ, но и информационных технологий вообще.

Двуязычность тезауруса даст возможность отечественным ученым и специалистам быстрее и эффективнее ориентироваться в мировой ситуации в данной области, окажет им помощь при написании и переводе статей. В ситуации еще не до конца преодоленного отставания российской КЛ от англоязычной и наличия некоторых существенных различий, сохранившихся от изолированного развития российской КЛ в советский период, составление такого тезауруса выявляет различия и сходства между понятиями, используемыми в отечественной и зарубежной науке, и позволяет вводить новые понятия и лингвистические термины, отсутствующие в русском языке [17].

В данной работе рассматривается подход к разработке русско-английского тезауруса по компьютерной лингвистике. Описывается общее строение тезауруса, структура тезаурусных статей и набор связей между терминами. Особое внимание уделяется проблеме выбора и представления отношений между терминами, а также поддержанию логической целостности терминологической системы тезауруса.

2. Структура русско-английского тезауруса по КЛ

Проектирование структуры двуязычного тезауруса по компьютерной лингвистике выполнялось на основе анализа существующих отечественных и международных стандартов [4,5,20-22], регламентирующих построение информационно-поисковых тезаурусов (ИПТ), а также на основе анализа и обобщения накопленного к этому времени опыта разработки тезаурусов ИНИОН [12], РуТез [11] и др.

Отечественные и международные стандарты определяют основные единицы, которые могут включаться в тезаурус, и набор отношений между ними, устанавливают правила сбора массива лексических единиц, формирования словника, построения словарных статей и оформления ИПТ.

По своему составу ИПТ подразделяют на тезаурусы, все единицы которых являются дескрипторами (или предпочтительными терминами), и тезаурусы, выделяющие среди своих единиц дескрипторы и аскрипторы (непредпочтительные термины). При этом дескрипторы могут использоваться при индексировании документов и в поисковых запросах, а аскрипторы (как текстовые входы) подлежат замене одним или несколькими дескрипторами [11].

В зависимости от языковой направленности тезаурусы разделяются на одноязычные и многоязычные.

Многоязычный информационно-поисковый тезаурус (МИПТ) содержит термины из нескольких естественных языков и представляет эквивалентные по смыслу понятия на каждом из них. В качестве основной структурной единицы МИПТ может рассматриваться составной дескриптор, собранный из эквивалентных дескрипторов одноязычных версий, связанных средствами для указания эквивалентности.

Построение русско-английского тезауруса по КЛ выполняется в соответствии с требованиями межгосударственного стандарта ГОСТ 7.24-2007 [4], который разработан с учетом основных нормативных положений международного стандарта ISO 5964-1985 [22] и устанавливает состав, структуру и основные требования к построению МИПТ. Русско-английский тезаурус по КЛ разрабатывается как набор одноязычных версий МИПТ, при этом выполняется согласованное построение одновременно двух версий тезауруса – русскоязычной и англоязычной.

Разработка каждой из одноязычных версий тезауруса выполняется на основе международного стандарта ISO 2788-1986 [21], межгосударственного стандарта ГОСТ 7.25-2001 [5] и американского стандарта Z39.19-2005 [20].

2.1 Структура словарной статьи

Основными единицами тезауруса являются термины предметной области (ПрО), которые разделяются на дескрипторы и аскрипторы. Согласно ГОСТ 7.25-2001, в ИПТ включаются следующие типы лексических единиц (ЛЕ): одиночные слова (существительные, прилагательные, глаголы, наречия), именные словосочетания, лексически значимые компоненты сложных слов, сокращения слов и словосочетаний[1]. ЛЕ объявляются эквивалентными в ИПТ, образуя класс эквивалентности, если замена одной ЛЕ на другую не приводит к изменению смысла текста, существенному для поиска информации. Одна из лексических единиц класса эквивалентности выбирается в качестве представителя этого класса и получает статус дескриптора, остальные ЛЕ получают статус аскриптора. При этом статус аскриптора получают также и термины, представляемые аббревиатурами или иными вариантами написания (через дефис, с пробелом и т.п.).

В состав словарной статьи тезауруса вне зависимости от статуса термина входят следующие элементы:

- *Название термина* предметной области, который представляет собой слово, словосочетание или лексически значимый компонент сложного слова естественного языка.
- *Язык*, на котором дано название термина.
- *Комментарий*, включающий правила и рекомендации использования термина, а также замечания и пояснения автора словарной статьи.
- *Автор словарной статьи* – задается для контроля процесса коллективной разработки тезауруса.

Термины-дескрипторы, кроме перечисленных выше атрибутов, описываются следующими дополнительными атрибутами:

- *Определение термина*, поясняющее на языке термина его смысл или значение. Наличие в тезаурусе определений терминов делает возможным его использование не только в качестве инструмента для ручного или автоматизированного индексирования, но и в качестве источника систематизированных знаний о данной ПрО.
- *Релятор*, представляющий собой помету, введенную для различения омонимичных терминов (омографов) в рамках описываемой ПрО. Он является частью термина и поясняет его значение, относя его к определенной понятийной категории или предметно-тематической области (в контексте данной работы – подобласти КЛ или смежной ей области/подобласти знаний). Например, для различения двух понятий, образованных на основе словосочетания РАЗМЕТКА ТЕКСТА, могут быть использованы реляторы ПРОЦЕСС и ОБЪЕКТ. В результате мы получаем два разных термина-дескриптора РАЗМЕТКА ТЕКСТА (ПРОЦЕСС) и РАЗМЕТКА ТЕКСТА (ОБЪЕКТ).
- *Область/подобласть знания*, к которой относится данный термин-дескриптор.
- *Признак корневого термина (Top Term)*, указывающий на то, что дескриптор находится на самом верхнем уровне какой-либо иерархии понятий.

Термины тезауруса связываются различными семантическими отношениями, отражающими место каждого термина в системе понятий выбранной ПрО.

Для связи дескрипторов с аскрипторами используются отношения синонимии. Так, если дескриптор может однозначно во всех контекстах заменить какой-то аскриптор, то он связывается с ним отношением «Синоним»; при этом также устанавливается обратное отношение от аскриптора к дескриптору – «Смотри». Для моделирования других соотношений между аскрипторами и дескрипторами в соответствии с ГОСТ 7.25-2001 в тезаурус вводятся отношения, позволяющие задавать связи между аскрипторами и альтернативными дескрипторами или представлять аскриптор комбинацией дескрипторов. Если нет однозначного соответствия между дескрипторами и аскрипторами, то используются отношения «Используй альтернативно» или «Используй комбинацию», задающие соответствие между аскрипторами и заменяемыми ими дескрипторами; при этом вводятся обратные им отношения «Сравни альтернативный выбор» и «Сравни комбинацию». Например, аскриптор ПАРТИЦИПАНТ можно связать отношением «Используй альтернативно» с дескрипторами СЕМАНТИЧЕСКАЯ ВАЛЕНТНОСТЬ и УЧАСТНИК СИТУАЦИИ. В то же время аскриптор СИСТЕМА СТАТИСТИЧЕСКОГО МАШИННОГО ПЕРЕВОДА может быть представлен с помощью связи «Используй комбинацию» как комбинация (сочетание) двух дескрипторов – СИСТЕМА МАШИННОГО ПЕРЕВОДА и СТАТИСТИЧЕСКИЙ МАШИННЫЙ ПЕРЕВОД.

Для отражения семантических связей между понятиями, выражаемыми дескрипторами, в одноязычных версиях устанавливаются иерархические и ассоциативные отношения.

Между дескрипторами вводятся такие иерархические отношения, как недифференцированная иерархическая связь «Выше», направленная от нижестоящего дескриптора к вышестоящему, родовидовая связь «ВышеРод», устанавливаемая между двумя дескрипторами, когда объем понятия нижестоящего дескриптора входит в объем понятия вышестоящего дескриптора, партонимическая связь «ВышеЦелое», задаваемая между двумя дескрипторами в том случае, когда нижестоящий дескриптор представляет компонент объекта, обозначаемого вышестоящим дескриптором. Вводятся также обратные им отношения: «Ниже», «НижеВид», «НижеЧасть».

Для задания отношений между дескрипторами, представляющими класс понятий и экземпляр этого класса, были выбраны связи «ВышеКлассЭкземпляра» и «НижеЭкземпляр».

При установлении иерархических отношений для некоторых дескрипторов можно указать признак «Аспект деления иерархии». Так, например, в иерархии, построенной по отношению «НижеВид», МАШИННЫЙ ПЕРЕВОД по признаку «подход» разделяется на СТАТИСТИЧЕСКИЙ МАШИННЫЙ ПЕРЕВОД, МАШИННЫЙ ПЕРЕВОД НА ОСНОВЕ ПРАВИЛ и МАШИННЫЙ ПЕРЕВОД, ОСНОВАННЫЙ НА ПРЕЦЕДЕНТАХ, а по признаку «степень участия человека» – на ПОЛНОСТЬЮ АВТОМАТИЧЕСКИЙ ПЕРЕВОД и ЧЕЛОВЕКО-МАШИННЫЙ ПЕРЕВОД.

Таким образом, один и тот же дескриптор одновременно может входить в несколько иерархий понятий, построенных по различным отношениям («Выше», «ВышеРод», «ВышеЦелое») и по различным аспектам деления иерархии.

Связи между дескрипторами, отличные от иерархических отношений и отношений синонимии, задаются отношением «Ассоциируется с». Такое отношение позволяет задавать произвольные ассоциативные связи между дескрипторами, например, отношения, выражающие зависимости вида «процесс-объект», «причина-следствие» и др.

Чтобы указать эквивалентность дескрипторов из разных одноязычных версий между ними устанавливается отношение «Эквивалент на другом языке». Если понятие не может быть выражено на другом языке одним дескриптором, тогда для него в соответствии с ГОСТ 7.24-2007 указывается в качестве эквивалента комбинация нескольких дескрипторов.

2.2 Представление источников терминов

Для подтверждения актуальности введенных в тезаурус терминов и ознакомления пользователей тезауруса с практикой их употребления для каждого термина задаются его связи с источниками, т.е. текстовыми документами или коллекциями текстовых документов, в которых данный термин встречается или определяется.

Этим целям служат два отношения: связь «Встречается в», при которой можно указать частоту встречаемости термина в источнике, если источник – коллекция текстов, и связь «Встречается в части документа», с помощью которой отмечается, что данный термин встречается в предметном указателе или глоссарии источника, что указывает на важность термина и повышает степень доверия к нему. Термины-дескрипторы, снабженные толкованиями-определениями, связываются с источником определения с помощью отношения «Дается определение в».

В тезаурусе источники описываются следующими параметрами: название, библиографическая ссылка, язык, тип (книга, монография, научная статья, документация, учебник, словарь, тезаурус, интернет-ресурс, коллекция текстов и др.), краткое описание и адрес в сети Интернет. Для коллекции текстов дополнительно задается число текстов и словоупотреблений.

3. Реализация электронной версии тезауруса

Для представления тезауруса в виде электронного ресурса было разработано двухуровневое хранилище данных, а для его разработки и сопровождения – редактор тезауруса.

На первом уровне хранятся структуры тезауруса, определяющие схемы тезаурусных статей, вид и свойства отношений, задаваемых между терминами, а также характеристики источников терминов и их определений. Создание и настройка структуры тезауруса осуществляется в специальном разделе редактора тезауруса. В частности, здесь определяются классы терминов и типы источников терминов, а также набор отношений и их свойства. Причем могут быть заданы не только структурные свойства отношений – путем указания типа их аргументов и задания ограничений на существование (число) и обязательность связей, но и формальные свойства – приписыванием отношениям математических свойств (симметричность, рефлексивность, транзитивность, асимметричность, антирефлексивность) и заданием для них обратных отношений.

Второй уровень обеспечивает хранение тезаурусных статей и описаний источников. Для задания терминов, их определений и источников, а также для установления связей между ними редактор тезауруса предоставляет экспертам-лингвистам удобный интерфейс. Заметим, что сразу после завершения ввода и/или редактирования описаний терминов, источников и связей между ними, новая информация становится доступной через пользовательский web-интерфейс тезауруса.

Редактор тезауруса реализован как web-приложение и доступен зарегистрированным пользователям через Internet. С целью обеспечения распределенной коллективной разработки в редакторе тезауруса поддерживается механизм делегирования прав экспертам разных уровней. В соответствии с этим механизмом только эксперты самого высокого уровня могут редактировать структуры тезауруса, а эксперты других уровней – только его содержание (описание терминов и источников). При этом действует следующее ограничение: два эксперта не могут одновременно редактировать одну и ту же словарную статью (или описание источника).

Кроме того, действует правило, по которому редактировать словарную статью может только ее автор. Если кто-то из экспертов захочет внести изменения в «чужую» статью, он может согласовать такую возможность с ее автором, в частности, через специальный форум, на который имеется ссылка в электронном тезаурусе.

Для того чтобы тезаурус мог использоваться при индексировании и поиске текстовых документов, он должен представлять целостную и непротиворечивую систему понятий ПрО. Это обеспечивается встроенными в редактор терминов механизмами вывода и поддержки логической целостности системы понятий тезауруса, работа которых базируется на описаниях свойств отношений тезауруса, представленных в редакторе тезауруса в виде аксиом и ограничений.

В частности, на основе этих свойств происходит корректное установление связей между терминами тезауруса, при необходимости осуществляется их автоматическое добавление и/или удаление. Кроме того, регулируются ограничения на существование и число тех или иных связей между терминами тезауруса в зависимости от их принадлежности к тем или иным классам.

Дескриптор	
название	машинный перевод
язык	русский
релятор	
определение 1	Область научных исследований, экспериментальных разработок и уже функционирующих систем, в которых к процессу перевода с одного естественного языка на другой привлекается ЭВМ.
определение 2	Процесс перевода текстов (письменных, а в идеале и устных) с одного естественного языка на другой с помощью специальной компьютерной программы. Так же называется направление научных исследований, связанных с построением подобных систем.
автор словарной статьи	Кононенко И.С.
комментарий	Выше – Приложения КЛ

Связи объекта

Дается определение в (SourceDef)

Источник	определение
Интернет энциклопедия «Википедия»	2
Справочник по искусственному интеллекту	1

Синоним (Syn)

Аскриптор
автоматический перевод
АП
МП

Эквивалент на другом языке (Trans)

Дескриптор
machine translation

Ниже (NT)

Дескриптор
автоматический перевод устной речи

Ниже вид (NTG)

Дескриптор	Аспект деления иерархии
машинный перевод на основе правил	подход
машинный перевод, основанный на прецедентах	подход
полностью автоматический перевод	участие человека
статистический машинный перевод	подход
человеко-машинный перевод	участие человека

Ниже часть (NTP)

Дескриптор
постредактирование (машинный перевод)
предредактирование (машинный перевод)
система машинного перевода

Встречается дескриптор в (SourceDescriptor)

Источник	частота
Коллекция текстов Диалог 2000-2010	318

Рис.1. Представление термина «Машинный перевод»

Например, если для рассмотренного в разделе 2.1. отношения «Смотри» задано обратное отношение («Синоним») и ограничение на существование связей («только одна связь данного типа для каждого термина-аскриптора»), то при связывании аскриптора АВТОМАТИЧЕСКИЙ ПЕРЕВОД и дескриптора МАШИННЫЙ ПЕРЕВОД отношением Смотри (АВТОМАТИЧЕСКИЙ ПЕРЕВОД, МАШИННЫЙ ПЕРЕВОД) произойдет создание обратной связи Синоним (МАШИННЫЙ ПЕРЕВОД, АВТОМАТИЧЕСКИЙ ПЕРЕВОД) (если таковой еще не существует), а также для аскриптора АВТОМАТИЧЕСКИЙ ПЕРЕВОД будет обеспечиваться запрет на создание связей «Смотри» и «Синоним» с другими дескрипторами.

Для обеспечения доступа к электронному тезаурусу был разработан пользовательский web-интерфейс, который представляет пользователю содержимое тезауруса в виде сети взаимосвязанных информационных объектов – элементов тезауруса: терминов (дескрипторов и аскрипторов) и описаний источников терминов и их определений. Набор атрибутов терминов и связей, установленных между ними, соответствует структуре тезауруса, описанной в разделе 2.1.

При навигации по тезаурусу обеспечивается возможность выбора необходимых пользователю терминов, детального просмотра их описаний (тезаурусных статей), а также описаний источников (публикаций или коллекций текстов), в которых встречается термин и/или его определение.

Пользователь может указать, какой тип информации его интересует – все термины, дескрипторы, аскрипторы или источники терминов. При этом ему выдается полный список имеющихся в тезаурусе объектов выбранного типа, который отображается в виде html-страницы, содержащей набор ссылок на эти объекты.

Информация о конкретном объекте и его связях также отображается в виде html-страницы (Рис.1). При этом объекты, связанные с данным объектом, представляются на его странице в виде гиперссылок, по которым можно перейти к их детальному описанию.

Дальнейшая навигация по тезаурусу представляет собой процесс перехода от одних объектов тезауруса к другим по заданным между ними связям, отражающим существующие между ними – тезаурусные (между терминами) или библиографические (между терминами и источниками) – отношения.

4. Методика выбора терминов для включения в тезаурус

Важным моментом при построении тезауруса является методика подбора терминов – кандидатов на включение в тезаурус, – а также выбор терминов-дескрипторов из множеств синонимичных терминов.

Выбор терминов для включения в русско-английский тезаурус по КЛ сопряжен с трудностями, которые обусловлены особенностями самой КЛ как новейшей науки и состоянием ее развития в России. Здесь важно отметить следующие факторы, характеризующие КЛ в целом и русскоязычную КЛ (РКЛ), в частности:

- междисциплинарный характер КЛ;
- неоднородность ПрО «Компьютерная лингвистика»;
- неравномерность развития отдельных направлений КЛ;
- отличие русскоязычной КЛ от англоязычной (в частности, отставание отдельных направлений РКЛ).

Ранее КЛ рассматривалась как часть исследовательского направления «искусственный интеллект» (ИИ). Терминология этого направления считается зрелой: «Специальная терминология по искусственному интеллекту и интеллектуальным системам начала формироваться в 60-е годы XX в. Первый этап формирования терминологии всегда отличается наличием многих синонимических терминов, которые используют различные школы и группы специалистов. На этом этапе термины быстро возникают и часть из них также быстро исчезает. К середине 70-х годов терминология в области искусственного интеллекта стала устанавливаться. Появились термины, которые признало подавляющее большинство специалистов. Все эти термины (за редким исключением) по происхождению англоязычные, так как именно в США проводились интенсивные исследования в этой области. Окончательно основная терминология закрепилась в первой половине 80-х годов» [18].

ИИ – это методологическая область, методы которой применимы к разным ПрО, в частности, активно применяются в КЛ в последнее десятилетие. Терминология КЛ в отдельных разделах продолжает сохранять черты первого этапа (наличие большого числа синонимов, например, в разделе семантических отношений). ИИ тоже считается междисциплинарной областью, однако по этому параметру ИИ и КЛ противоположны: ИИ междисциплинарна, потому что ее методы применяются в разных дисциплинах, КЛ – потому что она вбирает в себя разные дисциплины, такие как лингвистика (разделы, связанные с обработкой текстов и речи), психология, некоторые разделы ИИ.

Следствием указанных выше факторов является отсутствие русскоязычных учебных и лексикографических источников, достаточно полно отражающих структуру современной КЛ, в отличие от англоязычных источников, где она представлена детально и отчетливо.

Учитывая вышеперечисленные особенности КЛ и связанный с ними недостаток современной справочной русскоязычной литературы по КЛ, при разработке тезауруса использовались источники «живых» терминов РКЛ и их толкований, и именно они фиксируются в словарных статьях тезауруса.

В качестве основного источника русскоязычных терминов была выбрана коллекция текстов докладов, представленных на международной конференции «Диалог» в 2000-2010 гг., как «зеркала», отражающего термины РКЛ в их реальном употреблении.

К данной коллекции была применена словарная технология [16], с помощью которой на базе лингвистических моделей (морфологического и локального синтаксического анализа) и статистических показателей был создан список статистически значимых в данной ПрО слов и словосочетаний – кандидатов в термины ПрО. Затем этот список был обработан (отфильтрован) экспертами в области КЛ, которые существенно опирались не только на знания о предмете и направлениях КЛ, но и на общелингвистические представления о терминологичности и путях формирования терминологических словников. Таким образом, наш подход, учитывающий предварительное структурирование ПрО, согласуется с общей методикой формирования словников на базе классификационных схем предметных областей (см., например, [14]).

Для английской части словника, с учетом русско-английской направленности создаваемого тезауруса выбирались переводные эквиваленты из доступных англоязычных источников по КЛ.

С другой стороны, чтобы дополнить картину РКЛ в тех ее разделах, где имеются пробелы, при сборе терминов по таким разделам пришлось опираться преимущественно на англоязычные источники. Так, учитывая скачок, совершенный в течение последних нескольких лет в такой высокотехнологичной подобласти КЛ, как речевые технологии, а также тот факт, что это направление слабо представлено в коллекции «Диалог», при сборе терминов для этой подобласти была применена обратная методика, т.е. в качестве основных использовались англоязычные источники: предметные указатели нескольких современных и наиболее авторитетных англоязычных книжных источников обзорно-учебного профиля и глоссарии, входящие в документацию известных звуковых анализаторов. На данной терминологической базе был составлен англо-русский словник параллельных терминов.

Достаточно сложной оказалась и проблема выбора основного термина-дескриптора из множества синонимичных терминов. Прежде всего, эта проблема связана с появлением новых понятий и соответствующих им терминов. Так, появление систем *translation memory* в сфере автоматизированного перевода привело к широкому использованию практиками-переводчиками термина *память переводов*, который не был принят научным сообществом, противопоставившим ему термин *переводческая память* (синонимический ряд: *переводческая память* – 8, *память переводов* – 0, *архив переводов* – 1, *накопитель переводов* – 0, *копилка переводов* – 0)[2].

Развитие некоторых направлений КЛ (например, таких как *автоматический перевод в режиме онлайн*) приводит к столкновению вариантов старых терминов. Так, тезаурус ИНИОН [19] и ЛЭС [10] основным термином в паре *автоматический перевод* и *машинный перевод* считают *автоматический перевод*, присвоив ему статус дескриптора. Однако показатели встречаемости в коллекции «Диалог» говорят в пользу термина *машинный перевод*: *машинный перевод* – 318 vs. *автоматический перевод* – 58[3]. Интернет-энциклопедии «Википедия» и «Кругосвет», а также учебники придерживаются этой же традиции. На сайте Европейской ассоциации машинного перевода [15] также отмечается, что термин *machine translation*, хоть и звучит архаично, но, тем не менее, сохраняется как основной общий термин для всей области. В данном случае эксперты согласились с этой точкой зрения.

Таким образом, при выборе терминов дескрипторов мы опирались не только на статистику, но и на традиции словоупотребления, сложившиеся к настоящему времени в лингвистическом научном сообществе.

5. Заключение

В докладе представлен подход к разработке русско-английского электронного тезауруса по компьютерной лингвистике, общий состав и структура которого были разработаны на основе международных и отечественных стандартов.

При разработке программных компонентов электронной версии тезауруса (хранилища данных, пользовательского интерфейса и редактора) использовалась технология [7], которая была ранее применена для создания портала знаний по компьютерной лингвистике [3].

Хотя рассмотренные средства разрабатывались для создания русско-английского тезауруса по компьютерной лингвистике, благодаря наличию средств настройки структуры тезауруса и поддержки ее семантических свойств они могут быть использованы для построения многоязычных тезаурусов для любых языков и предметных областей.

В настоящее время ведется активная разработка тезаурусных статей и заполнение ими контента электронного тезауруса, который на данный момент включает более 1000 терминов КЛ, около 3500 связей между терминами и более 120 источников терминов и их определений.

Литература

[1] Ахманова О.С. Словарь лингвистических терминов. – 3-е изд., стер. – М.: УРСС, 2005. – 576 с.

[2] Большой энциклопедический словарь (БСЭ) / гл. ред. А.М.Прохоров. - Изд. 2-е, перераб. и доп. – М. : Большая Российская энциклопедия; – – СПб.: Норинт, 2004. – 1456 с.

[3] Боровикова О.И., Загорулько Ю.А., Загорулько Г.Б., Кононенко И.С., Соколова Е.Г. Разработка портала знаний по компьютерной лингвистике // Труды 11-ой национальной конференции по искусственному интеллекту с международным участием КИИ-2008. – М.: ЛЕНАНД, 2008. –Т.3. –С.380-388.

[4] ГОСТ 7.24-2007. Система стандартов по информации, библиотечному и издательскому делу. Тезаурус информационно-поисковый многоязычный. Состав, структура и основные требования к построению. (Введен в действие с 1 июля 2008 г.).

[5] ГОСТ 7.25-2001. Система стандартов по информации, библиотечному и издательскому делу. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления. (введен в действие с 1 июля 2002 г.)

[6] Демьянков В.З. Англо-русские термины по прикладной лингвистике и автоматической переработке текста. Вып. 2. Методы анализа текста // Тетради новых терминов. № 39. – М.: ВЦП, 1982.

[7] Загорулько Ю.А., Боровикова О.И. Подход к построению порталов научных знаний // Автометрия. Новосибирск: 2008. Т. 44. № 1. С. 100–110.
[8] Интернет-энциклопедия «Википедия» http://ru.wikipedia.org
[9] Никитина С.Е. Тезаурус по теоретической и прикладной лингвистике. – М.: Наука, 1978.
[10] Лингвистический энциклопедический словарь. //Под ред. В. Н. Ярцевой. М.:Советская энциклопедия, 1990. — 685 с. [3 изд. 2002.]
[11] Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Издательство Московского университета, 2011. – 512 с..
[12] Мдивани Р.Р. О разработке серии тезаурусов по социальным и гуманитарным наукам // НТИ, сер. 2, №7, 2004. с. 1-9.
[13] Онлайн Энциклопедия «Кругосвет»: [сайт]. [2001-2009]. URL: http://www.krugosvet.ru/
[14] Перерва В.М. О принципах и проблемах отбора терминов и составления словника терминологических словарей // Проблематика определений терминов в словарях разных типов. – Л., 1976. – С. 190-204.
[15] Веб-сайт EAMT (The European Association for Machine Translation). http://www.eamt.org/
[16] Сидорова Е.А. Многоцелевая словарная подсистема извлечения предметной лексики // Труды международной конференции Диалог' 2008 «Компьютерная лингвистика и интеллектуальные технологии». М.: РГГУ, 2008. Вып. 7 (14). –С. 475-481.
[17] Соколова Е.Г., Семенова С.Ю., Кононенко И.С., Загорулько Ю.А., Кривнова О.Ф., Захаров В.П. Особенности подготовки терминов для русско-английского тезауруса по компьютерной лингвистике // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог» (Бекасово, 25-29 мая 2011 г.). Вып. 10(17). –М.: РГГУ, 2011. –С.644–655.
[18] Толковый словарь по искусственному интеллекту / Авторы-составители: А.Н. Аверкин, М.Г. Гаазе-Рапопорт, Д.А. Поспелов. – М.: Радио и связь, 1992. –256с. (http://www.raai.org/library/tolk/aivoc.html)
[19] Языкознание. Информационно-поисковый тезаурус ИНИОН РАН. – М., 2007.
[20] ANSI/NISO Z39.19-2005 Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies (Periodic Review).
[21] ISO 2788-1986. Documentation – Guidelines for the establishment and development of monolingual thesauri. Ed. 2.
[22] ISO 5964-1985. Documentation - Guidelines for the establishment and development of multilingual thesauri, IDT (Revised by: ISO/DIS 25964-1 Under development).

Approach to Development of Russian-English Thesaurus on Computational Linguistics

© Yu.A. Zagorulko, O.I. Borovikova, I.S. Kononenko, E.G. Sokolova

The paper presents an approach to development of Russian-English thesaurus on Computational Linguistics. A general structure of the thesaurus, composition of the thesaurus entries and set of relations between terms of the thesaurus are described. The problems of choice of terms for inclusion in the thesaurus and the preferred terms (descriptors) from set of synonymous terms are discussed.

Features of implementation of online version of the thesaurus are outlined. The paper gives a particular attention to maintenance of a logical consistency of the thesaurus terminology system and to providing a convenient access to the thesaurus content.

♣ Работа выполнена при финансовой поддержке РГНФ (проект № 10-04-12108в).

[1] На начальном этапе мы включаем в тезаурус только существительные и именные словосочетания.

[2] Здесь приводятся частотные характеристики терминов в коллекции «Диалог»

[3] Поиск в Интернете дает обратное соотношение: *машинный перевод* – 640000, *автоматический перевод* – 1960000, которое объясняется тем, что если речь идет о МП с языка на язык (а не о переводе на другой тариф и т.п.), основную часть ответов составляет реклама онлайн-переводчиков, т.е. имеется в виду разновидность полностью автоматического перевода (онлайн-перевод).