Ю. А. Киселев, М. Ю. Мухин, С. В. Поршнев

РАЗРАБОТКА ЭЛЕКТРОННЫХ ТЕЗАУРУСОВ РУССКОГО ЯЗЫКА: ОПЫТ МЕЖДИСЦИПЛИНАРНОГО СОТРУДНИЧЕСТВА

УДК 81`33:51-7

Авторы публикации рассматривают проблемы междисциплинарного взаимодействия представителей гуманитарного и естественнонаучного знания на примере создания электронного тезауруса русского языка. Приводятся примеры дискуссионных случаев, связанных с формализацией семантических связей между словами. В ходе работы над конкретным проектом возможны компромиссные решения в области моделирования языковых данных и взаимовлияние специалистов различных предметных областей.

Ключевые слова: электронный тезаурус, словарь, лексикография, семантика, междисциплинарное взаимодействие, синонимы, антонимы.

1. Лингвистика и информационные технологии. Лингвистические ресурсы (корпуса, тезаурусы, поисковые системы, машинный перевод) как сфера междисциплинарного взаимодействия.

Рубеж XX–XXI вв. стал периодом кардинально быстрого развития междисциплинарных направлений в науке. Одним из заметных примеров научно-практического взаимодействия стало вовлечение лингвистики, с ее фундаментальными знаниями о языке, в развитие современных информационных технологий. Так, многие пользователи поисковых систем никогда не слышали о том, что не только их запросы, но и индексируемый контент интернета проходит через автоматические лингвистические анализаторы. Наиболее очевидным и необходимым для

языков с развитым словоизменением является морфологический анализ. Например, если пользователю Интернета необходимо найти компанию, устанавливающую пластиковые окна, ему не нужно печатать в строку поиска все формы слова *окно*: *окно, окна, окну, окном, окне* и т. д. Анализирует и синтезирует грамматические формы слов программа морфологического анализа. Задумаемся о том, что типичное русское существительное имеет 12 форм, прилагательное может иметь более ста, а глагол – более ста пятидесяти форм.

В XX в. стало очевидно, что без лингвистических компонентов невозможно развитие машинного перевода, алгоритмов проверки грамотности, авторизации текста и – в целом – элементов искусственного интеллекта.

2. Электронные тезаурусы. Проблема создания электронного тезауруса русского языка и проект YARN (Yet Another RussNet).

Одним из междисциплинарных лингвистических ресурсов является электронный тезаурус. Вообще, тезаурус – это словарь, в котором слова сгруппированы не по алфавиту, а по близости значений. В итоге рядом, в одной лексической группе, оказываются однородные ряды понятий. Если, например, названия месяцев (*август* и *январь*) в обычном алфавитном толковом словаре можно найти разных местах, то в тезаурусе они войдут в одну семантическую группу. Аналогично в одних группах тезауруса содержатся прилагательные цвета, глаголы движения и т. д. [см., например: Большой толковый словарь русских глаголов; Словарь-тезаурус современной русской идиоматики; Большой толковый словарь русских существительных; Словарь-тезаурус русских прилагательных; Русский семантический словарь].

Сегодня тезаурусы незаменимы в практике преподавания и изучения языков. Активно они используются для решения проблем семантического анализа в интеллектуальных системах, информационном поиске, машинном переводе и т. д. [подробнее см.: Киселев, Мухин, Поршнев,

2015; Киселев, Мухин, Поршнев, 2018]. В то же время полного электронного тезауруса русского языка до сих пор не существует. На фоне ряда продолжающихся или заброшенных проектов возникла идея принципиально нового лексикографического ресурса, основанного на краудсорсинге, т.е. приложении усилий волонтеров, которые пополняют и редактируют тезаурус, вводят туда новые слова и значения и связывают их различными семантическими отношениями. В создании и развитии этого ресурса приняли участие ученые из Екатеринбурга (УрФУ, СКБ-Контур), Москвы и др. городов. Проект получил название YARN (Yet Another RussNet), а его создание было поддержано грантом РГНФ [см.: Браславский, Мухин, Ляшевская, Бонч-Осмоловская, Крижановский, Егоров; YARN].

Естественно, что любой такой современный проект требует вложения усилий специалистов в областях гуманитарного и естественнонаучного знания – в первую очередь лингвистов и программистов. В таком коллективе лингвист не может оставаться отвлеченным гуманитарием-теоретиком и в итоге ориентируется на создание более строгих языковых моделей. Программист со своей стороны тоже не просто технический работник, он принимает содержательное участие в оценке и формализации гуманитарных сущностей.

3. Семантические отношения и база данных тезауруса. Проблема разногласий в отношении формализации языковых данных.

Даже в самом сбалансированном междисциплинарном коллективе при совместной разработке языкового ресурса между представителями гуманитарного и естественнонаучного знания возникают специфические расхождения. Приведем два примера.

(1) Тезаурус призван отражать известные типы смысловых отношений между словами – синонимия, антонимия, гипо-гиперонимия (т. е. родо-видовые отношения) и ряд других. При планировании архитектуры базы данных программист логично, со своей точки зрения, предполагает, что,

если ряд слов-синонимов (*Нагреваться, греться, теплеть, раскаляться*…) сосредоточен вокруг одного понятия, а другой (*Охлаждаться, остывать, стынуть, холоднеть*…) – вокруг противоположного, то единая семантическая связь антонимии может соединять целые ряды, а не конкретные слова. Однако система языка сопротивляется такой установке (сомнительными, например, здесь будут антонимы *греться – холоднеть*). Выделение антонимов традиционно является семантически гораздо более строгой операцией, чем составление синонимических рядов.

Можно сказать даже, что перед нами две принципиально разные интуитивные и лексикографические установки. В синонимические ряды намеренно включают слова, имеющие значительные семантические и стилистические различия. Словари синонимов и создаются, в частности, для того, чтобы подчеркнуть эту разницу. Сравним синонимы *пьяный* и (разговорно-сниженные) *готовый, готовенький*. И стилистические, и семантические (оттенок меры и степени), и грамматические (употребление слова *готовый* преимущественно в краткой форме – *готов*) различия не мешают появлению этих слов в ряду синонимов. Однако в традиционных словарях антонимов мы не найдем пары типа *готовый* (*готовенький*) – *трезвый*, даже в варианте *готов – трезв*.

Самую яркую, бесспорную антонимию формируют основные значения слов. Синонимы с семантическими различиями обычно можно воспринять и вне контекста, а для антонимов такого рода всегда необходимо пояснение: *нос* (передняя часть чего-либо) – *хвост*, *яркий – рядовой* (т. е. обычный), *популярный – специальный* (об изложении), *плоский* (неумный, пошлый) – *остроумный, тонкий* и т. п.

В итоге отношения антонимии в базе данных приходится прописывать не между группами слов, а между конкретными словами.

(2) Термин «синонимы» можно понимать расширенно – как слова одной тематической группы, или когипонимы. Сравним: *пальто – шинель, кепка – бейсболка – фуражка, дума – сейм – рада – парламент, латы –*

броня – кольчуга, царь – император – хан – шах – падишах. Такие слова демонстрируют различные функциональные и национально-культурные различия, и с традиционной точки зрения синонимами они в основном не являются. Однако потребности автоматической обработки текста сегодня пока что предполагают более «грубый», менее гранулярный принцип классификации для более эффективной работы программ семантического анализа. Каждый такой случай нуждается в отдельном рассмотрении и точечном решении. В итоге некоторые тематические ряды все же входят в электронный тезаурус как условно понимаемые синонимы.

Таким образом, в ходе общей работы по планированию архитектуры базы данных словаря и обсуждению готового словарного контента обнаруживается очевидное междисциплинарное взаимодействие. Лингвист, решая классификационные задачи, в разных случаях может согласиться с некоторым «упрощением» модели языковой системы в соответствии с прикладными потребностями. Программист в свою очередь понимает необходимость усложнения архитектуры базы данных словаря, чтобы более гибко отразить языковые особенности.

Между естественниками и гуманитариями существует известная вневременная дискуссия. Пафос первых иногда сводится к критике расплывчатости гуманитарных объектов и определений. Пафос вторых – к указанию на излишний прагматизм оппонентов и их ошибочную уверенность в заведомом понимании этих гуманитарных объектов. Например, это касается непрофессиональных суждений об устройстве языка как родного с детства материала.

В процессе работы над конкретным проектом, каким является электронный тезаурус, эта общая по характеру дискуссия перестает быть актуальной, а на первый план выходит индуктивное решение вполне конкретных прикладных лингвистических проблем. Именно в этом случае опыт междисциплинарного сотрудничества становится по-настоящему

ценным, обогащает его участников и приводит к созданию общественно значимого интеллектуального ресурса.

_____

Браславский П. И., Мухин М. Ю., Ляшевская О. Н., Бонч-Осмоловская А. А., Крижановский А. А., Егоров П. YARN: начало // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог». М., 2013 // Dialog-21.ru: сайт [Электронный ресурс]. URL: http://www.dialog-21.ru/digests/dialog2013/materials/pdf/BraslavskiyP_YARN.pdf (дата обращения 10.10.2018).

Большой толковый словарь русских глаголов: Идеографическое описание. Синонимы. Антонимы. Английские эквиваленты / под ред. Л. Г. Бабенко. М., 2007.

Большой толковый словарь русских существительных: Идеографическое описание. Синонимы. Антонимы / Под общ. ред. проф. Л. Г. Бабенко. М., 2008.

Киселев Ю. А., Мухин М. Ю., Поршнев С. В. Автоматизированные методы выявления семантических отношений для электронных тезаурусов. Монография. М., 2018.

Киселев Ю. А., Мухин М. Ю., Поршнев С. В. Современное состояние электронных тезаурусов русского языка: качество, полнота и доступность // Программная инженерия. 2015. № 6. С. 34–40 [Электронный ресурс]. URL: http://novtex.ru/prin/full/06_2015.pdf (дата обращения 10.10.2018).

Русский семантический словарь / под общей ред. Н. Ю. Шведовой [Электронный ресурс]. URL: http:// www.slovari.ru/default.aspx?s=0&p=235 (дата обращения 10.10.2018).

Словарь-тезаурус русских прилагательных. Распределение по тематическим группам / Под общ. ред. проф. Л. Г. Бабенко. М., 2016.

Словарь-тезаурус современной русской идиоматики / под ред. А. Н. Баранова, Д. О. Добровольского. М., 2007.

YARN (Yet Another RussNet) . Проект создания нового электронного тезауруса русского языка [Электронный ресурс]. URL: https://russianword.net/ (дата обращения 10.10.2018).

А. А. Керимов

ПОЛИТИЧЕСКОЕ ОБРАЗОВАНИЕ КАК ФАКТОР ФОРМИРОВАНИЯ ГРАЖДАНСКОЙ ИДЕНТИЧНОСТИ

УДК 323.212

Формирование гражданской идентичности является одним из важнейших условий успешного развития страны. Политическое образование может дать гражданину понимание задач развития страны и его места в их решении, помочь человеку осознать ценность гражданской идентичности. Через политическое образование можно выстроить систему целенаправленного формирования позитивной гражданской идентичности без культивирования исторических обид и негативной памяти, столь необходимой для создания благоприятного социального климата и устойчивого развития общества.

Ключевые слова: гражданин, гражданская идентичность, политологическое образование, патриотическое воспитание, социализация.

Формирование гражданской идентичности считается одной из приоритетных задач в развитии системы образования Российской Федерации. Очевидно, что «позитивная гражданская идентичность является ключевым фактором**,** определяющим вектор развития страны, поскольку она задает совокупность характеристик, от которых зависит устремленность нации к успеху» [Семененко, с. 3].