Рычкова, Л.В. Языковые ресурсы: традиции и инновации / Л.В. Рычкова // V Международная научно-практическая конференция «Прикладная лингвистика в науке и образовании» памяти Р.Г. Пиотровского (1922-2009). Материалы (Санкт-Петербург, 25 – 26 марта 2010) / Учебно-методическое объединение по направлениям педагогического образования, Российская академия естественных наук, Национальная ассоциация прикладной лингвистики, Ассоциация прикладной лингвистики СПБ, Российский государственный педагогический университет им. А.И. Герцена, Санкт-Петербургский государственный университет. – С.Петербург: Изд-во «Лема», 2010. – С. 306 – 311.

Рычкова Л.В.

Гродненский государственный университет

ЯЗЫКОВЫЕ РЕСУРСЫ: ТРАДИЦИИ И ИННОВАЦИИ

Любой текст, независимо от формы существования, является информационным ресурсом, так как отражает определенное, прежде всего, внелингвистическое и экстраязыковое знание. Рассматриваемый как информационный ресурс, текст определяется как принадлежащий к определенному стилю и тематике либо к определенному подъязыку / предметному языку / языку для специальных целей. Определение такой принадлежности чрезвычайно важно для снижения неоднозначности, свойственной любому тексту, и облегчает его обработку, в т.ч. все виды автоматизированной / автоматической обработки (индексирование для целей информационного поиска, другие виды компрессии, перевод). Рассматриваемый как языковой материал, текст также несет информацию, но информацию чисто языковую, отражающую особенности функционирования языковых / речевых единиц / объектов. Совокупность всех текстов на определенном языке составляет генеральную совокупность языкового материала. Наиболее важными признаками дифференциации языкового материала являются *аутентичность / неаутентичность* и *первичность / производность*. Любой *языковой материал порождается* в процессе речевой деятельности и может рассматриваться как ее результат. Что касается *языковых ресурсов*, то они всегда *создаются* (составляются) с определенной(ыми) целью(ями) на основе исходного языкового материала. Следовательно, любые языковые ресурсы могут рассматриваться как *производный* языковой материал. Аутентичность языкового материала – понятие достаточно сложное. С одной стороны, оно связано со спецификой порождения первичной / вторичной, монолингвальной / билингвальной / плюралингвальной личностью-адресантом, а с другой, - с прагматикой пользователя-адресата, с соответствием ожиданиям "потребителя".

Традиционными языковыми ресурсами являются "бумажные" словари и грамматики, используемые далеко не только лингвистами. Так, энциклопедические и научно-технические (в том числе терминологические) словарные продукты, моделируя определенную

предметную область, отражают специальные тексты как информационный ресурс. Лингвистические словари отражают инвентарь языка и его парадигматику. Грамматики – различные аспекты синтагматики, в том числе парадигмальную синтагматику.

Современные компьютерные технологии позволяют представлять любые тексты на любом языке/языках в устной и письменной форме, поэтому данные признаки можно считать "шумовыми" [Марусенко 1996]. Большинство электронных текстов являются вторичными, хотя, в силу развития социальных сетевых проектов и формирования многочисленных Интернет-сообществ, стремительно растет первичный "сетевой" языковой материал, отражающий как разговорную речь, так и многочисленные жанры "нете-/сетературы". Авторы статьи о прикладной лингвистике в электронной энциклопедии "Кругосвет" (www.krugosvet.ru), отмечая, что современная прикладная лингвистика "почти столь же многообразна, как и области практической деятельности человека", "к наиболее перспективным прикладным областям", в первую очередь, относят "гипертекстовые технологии, непосредственно связанные с эксплуатацией и развитием глобальной компьютерной сети Internet", а также "компьютерный дизайн текста и его компонентов, в том числе шрифта". Шрифты, в том числе электронные, - один из традиционных видов языковых ресурсов. К традиционным видам языковых ресурсов по праву можно отнести электронные версии (пусть даже усовершенствованные) традиционных словарей и грамматик, а также собственно машинные словари, создающие основу лингвистического обеспечения различного рода "интеллектуальных систем" (изменения, которые претерпели машинные словари в последнее время, обусловлены лишь совершенствованием технического обеспечения и не затронули сути самого этого вида языковых ресурсов).

Развитие компьютерных технологий порождает новые, инновационные, виды информационных и языковых ресурсов, среди которых наиболее важным видом ресурсов, позволяющих сочетать обе функции текстов, являются генеральные (национальные) корпуса текстов, зачастую рассматриваемые и выступающие в качестве основы для создания других видов специализированных языковых ресурсов. Сферу компьютерных языковых ресурсов характеризует бурный рост, с одной стороны, и стремление к их стандартизации, с другой. Стандартизация затрагивает не только процессуальные аспекты разработки и создания языковых ресурсов, но и исходный языковой материал, положенный в их основу, - тот набор идиомов, который находит репрезентативное отражение в электронных ресурсах и может быть идентифицирован при их обработке и выдаче результатов потребителю. Таким образом, достоверность получаемых данных напрямую связана с аутентичностью исходного языкового материала, которая может быть установлена только потребителем, для чего ему необходим доступ к метаданным, отражающим особенности исходного вида языкового

материала. С этой точки зрения, стремление опираться на материалы Интернета как лингвистический корпус имеет свои ограничения по сравнению со специально создаваемыми корпусами, особенно национальными мегаресурсами, обусловленные не только проблемой репрезентативности (понимаемой как адекватное отражение генеральной совокупности языкового материала), но и, в большей мере, сложностями определения степени достоверности получаемых данных. Справедливости ради, следует отметить, что в наиболее удачных проектах, таких, например, как Web as Corpus – Веб как корпус (http://webascorpus.org/) или KWICFinder - Поиск ключевых слов в контексте, то есть построение конкордансов на основе материалов Веба (http://kwicfinder.com/index.html), эти сложности до некоторой степени снимаются.

Разнообразие и экспоненциальный рост инновационных языковых ресурсов потребовали разработки принципов их менеджмента, для чего в рамках технической комиссии Международной организации стандартизации (International Standards Organization – ISO) TC37 "Терминология и другие языковые и содержательные ресурсы" (Terminology and Other Language and Content Resources), помимо трех подкомиссий – SC1: Принципы и методы (Principles and Methods), SC2: Терминография и лексикография (Terminography and Lexicography), SC3: Компьютерные приложения для терминологии (Computer Applications for Terminology), – была создана подкомиссия SC4: Менеджмент языковых ресурсов (Language Resources Management). Терминология, представляя собой языковой ресурс, напрямую соотносится с инженерной лингвистикой, ее методами и инструментарием. Так, инженерная лингвистика разрабатывает приемы и методики для записи, хранения и обработки языковых ресурсов, поэтому разработка специальных стандартов для сферы менеджмента языковых ресурсов на основе использования компьютерных технологий напрямую касается и терминологии. Одной из задач подкомиссии SC4 является стандартизация дефиниций и терминов, которые представляют базовые концепты сферы менеджмента языковых ресурсов, включая создание различных видов языковых ресурсов, их оценку и дальнейшую обработку. В рамках решения этой задачи была проведена инвентаризация целевой терминологии и частотное упорядочение терминов. Наиболее частотным термином в этом ряду оказался термин *annotation* - *разметка* со значением абсолютной частоты 562 (для сравнения: следующий в ряду термин – *WordNet* - имеет частоту 249). К области разметки можно отнести также следующие термины из списка (в скобках даются показатели частоты): *annotator* (139 – третья позиция в списке), *element* (130 – четвертая позиция в списке), *feature* (109 – пятая позиция в списке), *collocation* (91 – шестая позиция в списке), *annotation scheme* (71 – восьмая позиция в списке), *attribute* (59 – десятая позиция в списке), *annotated* corpus (47 – двенадцатая позиция в списке), *annotation tool* (40 – двадцать первая позиция в списке),

annotation process (38 – двадцать вторая позиция в списке), *annotation task* (26 – тридцать седьмая позиция в списке) и т.д. Таким образом, создание инновационных языковых ресурсов и их менеджмент напрямую зависит от принятой системы разметки. Не случайно каждый из ныне действующих национальных корпусов сопровождается подробным описанием используемых в нем видов лингвистической разметки, а инструкции поиска в корпусе напрямую отражают возможности получения данных в зависимости от особенностей размеченного корпуса. Действительно, именно от принятой системы разметки зависит, будет ли положенный в основу языкового ресурса корпус языкового материала рассматриваться как собственно информационный ресурс (в любом случае, он представляет собой информационный массив либо вербализованной информации, либо фактических сведений о языке / речи, тексте / дискурсе и т.п.). Двойственной природой текста и множественностью задач его обработки объясняется отсутствие (пока) единого стандарта менеджмента языковых ресурсов, а также стандартов создания различных типов языковых ресурсов. Существует несколько хорошо известных инициатив / стандартов "кодирования" (индексирования или разметки) текстов: Инициатива кодирования текста – Text Encoding Initiative (TEI) (www.tei-c.org), Стандарт кодирования текста – the Corpus Encoding Standard (CES / XCES) (www.xml-ces.org), Стандарты экспертной группы советников по инженерной лингвистике – the Expert Advisory Group on Language Engineering Standards (EAGLES) (www.ilc.cnr.it/EAGLES96/home.html), Международный стандарт по инженерной лингвистике – the International Standard for Language Engineering (ISLE) (www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm) и др. Разработано также достаточно большое количество программных средств ("менеджеров" языковых ресурсов, например, корпус-менеджеров), предназначенных для создания, аннотирования (разметки) и использования языковых ресурсов: MULTEXT (Multilingual Text Tools and Corpora) – Мультилингвальные текстовые инструменты и корпусы (www.lpl.univ-aix.fr/projects/multext); разработки Группы языковых технологий (the Language Technology Group) - LT XML (www.ltg.ed.ac.uk/software/xml), NITE (www.dfki.de/nite/main.html) и др.; GATE (General Architecture for Text Engineering) – Общая архитектура для обработки текста (http://gate.ac.uk/); ATLAS (Architecture and Tools for Linguistic Analysis Systems) – Архитектура и инструменты для систем лингвистического анализа (www.nist.gov/speech/atlas/) и др. Выбор стандарта и / или программы-менеджера (могут рассматриваться как лингвистические метаресурсы, или метаязыковые ресурсы) обеспечивает совместимость создаваемых при их помощи ресурсов и, соответственно, формирует определенный "тип" ресурсов по таким показателям, как структура ресурса в целом и структура самих данных, включая формат их представления и типы поисковых

объектов, задаваемые системой индексирования / разметки. Так, TEI, CES / XCES, MATE / NITE ориентированы на представление первичных данных и разметку, а стандарты по инженерной лингвистике (такие, например, как EAGLES или ISLE) направлены на извлечение смысла, то есть на задачи АОТ (автоматической обработки текста) – NLP (Natural Language Processing) в его традиционном и новейшем понимании компьютерной лингвистикой. ISLE, наряду с такими специальными метаязыковыми ресурсами, как OLIF (The open XML language data standard) – Взаимозаменяемый формат для открытого лексикона (www.olif.net/) и SALT (Standards-based Access to Lexicographical & Terminological multilingual resources) – Основанный на стандартах доступ к лексикографическим и терминологическим мультилингвальным ресурсам (www.loria.fr/projects/SALT/), направлен также на представление лексических (в том числе, и терминологических) данных. Особый интерес представляют ресурсы, ориентированные на представление знаний, такие, например, как RDF / OWL (www.w3.org/2004/01/sws-pressrelease.html.en) или Тематические карты - Topic Maps (www.topicmaps.org/). Существует также множество инициатив, направленных на решение задач представления метаданных, например, Инициатива открытых архивов – the Open Archives Initiative (www.openarchives.org), Инициатива метаданных ISLE (the ISLE Meta Data Initiative) – IMDI (www.mpi.nl/IMDI/), Сообщество открытых языковых архивов (Open Language Archives Community) - OLAC (www.language-archives.org). Такого рода ресурсы не могут рассматриваться только как языковые или лингвистические, они являются мультимодальными и многоцелевыми.

Особое место в ряду мультимодальных ресурсов занимают те из них, которые направлены на представление номенклатуры: товарных знаков, наименований продукции, свойств (качеств) различных видов промышленных и / или коммерческих продуктов и т.д. Так, стандартизация свойств / состояний продукции на различных этапах промышленного цикла ее создания чрезвычайно важна не только для интернациональных корпораций, осуществляющих глобальное / международное распределение труда, но и при эксплуатации импортного оборудования субъектами хозяйствования любой страны. По инициативе Института стандартизации Германии (DIN) в опоре на стандарты ISO 13584 и IEC 61360 интеллектуальной фирмой Paradine был разработан DIN Properties Dictionary (www.dinsml.net/opencms/opencms/index_de.html?_locale=en) – он-лайн ресурс, отражающий стандартизованную номенклатуру свойств промышленной продукции, по сути дела, представляющий собой специальную лексикографическую базу данных. Развитие е-бизнеса потребовало создания таких мультимодальных ресурсов, как каталоги электронных продуктов (е-каталоги). Примером такого каталога может служить BMEcat (European e-

business catalog standard) – Европейский каталог е-бизнеса, имеющий силу стандарта (www.bmecat.org).

В качестве примера своеобразной программы-менеджера такого рода ресурсов можно привести классифицирующую систему eCl@ss (www.eclass-online.com/) - рассматривается как международный стандарт для классификации и описания продуктов и услуг, направленный на поддержку малого и среднего бизнеса, в том числе в международном масштабе.

Как отмечает Рейнхард Вейссингер [Weissinger 2007], в последние годы наблюдается быстрый рост баз данных, используемых различными комиссиями ISO в целях менеджмента знаний, закрепленных в стандартах, путем структуризации их содержания. Примерами таких структурированных продуктов являются термины и дефиниции, графические символы, коды всех типов, словари данных, свойства продукции, элементы систем классификации и т.п. Важной тенденцией, находящей повсеместное отражение в стандартах ISO, особенно разрабатываемых в рамках TC37, является учет особенностей максимального количества языков и их вариантов. Так, традиционный для европейских языков алфавитный принцип упорядочения статей в лексикографических продуктах или подход к сегментации текста, основанный на пробелах между словоупотреблениями, " не работают " в таких языках, как китайский, корейский и многие другие. Принцип языкового разнообразия естественен при рассмотрении каждого идиома как ресурса и способствует привлечению внимания широкого круга специалистов к потенциалу идеологии "языка как ресурса", составляющей действенную альтернативу все еще доминантной в языковой политике многих стран идеологии "язык как проблема" или "язык как право" [Hornberger 2002].

Литература

Марусенко М.А. Атрибуция анонимных и псевдонимных текстов методами прикладной лингвистики // Прикладное языкознание. - СПб, 1996, с.469-473.

Hornberger N.H. Multilingual Language Policies and the Continua of Biliteracy: an Ecological Approach // Language Policy, 2002, N. 1, p. 27-51.

Weissinger R. Standards as databases and the development of knowledge // ISO Focus, November 2007, p. 36-37.