

Л.Н. Беляева

Российский государственный педагогический университет
им. А.И. Герцена, Санкт-Петербург

СЕТЕВЫЕ ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ В ПОДГОТОВКЕ ФИЛОЛОГА

В статье рассматривается научный и образовательный потенциал сетевых лингвистических ресурсов, обеспечивающих оперативное извлечение информации и данных, необходимое при подготовке современного филолога.

Ключевые слова: *сетевые лингвистические ресурсы, терминологические базы данных, извлечение информации, извлечение и перевод терминов.*

В современном мире экономики знаний именно деятельность филолога является базой для извлечения и анализа эмпирической информации, содержащейся в текстах на естественном языке [1–4]. Современным средством поддержки и сопровождения работы любого филолога являются информационные технологии (ИТ) и созданные на их основе лингвистические ресурсы. Однако часто именно филолог привычно воспринимает ИТ как реальную угрозу своему существованию в профессии, поскольку ему либо вообще не известны, либо недостаточно известны ни потенциал этих технологий в целом, ни конкретные возможности их применения для решения собственных задач – лингвистических, литературоведческих и методических. Незнание этих возможностей (или, что еще хуже, знание неполное и/или некорректное) приводит к тому, что филолог не владеет информацией о существующих в Сети и доступных лингвистических ресурсах, национальных корпусах текстов, словарных базах и т.п. (ср., например, такую справочную информацию в [5, 6]). Естественно, он не в состоянии выбрать нужные именно ему сервисы и системы и, следовательно, не способен их адекватно использовать.

В самом общем виде под сетевыми ресурсами принято понимать любые данные, к которым можно получить доступ при подключении к системе Интернет. Сетевые ресурсы могут по разным критериям объединяться в информационные сегменты, которые и являются основой для поддержки и сопровождения профессиональных образовательных пространств, формируя высокотехнологичную образовательную среду. Сетевые ресурсы сегодня представляют собой широко формируемую, разнообразную и доступную информацию. Основной проблемой при формировании информационного сегмента в ходе решения задач обучения и/или исследования является выбор тех из них, информацией которых можно пользоваться с полным доверием. При

формировании образовательного пространства филологии решение этой задачи и формирование принципов отбора рекомендуемых ресурсов становится принципиально важным [7, 8].

При этом следует учитывать, что образовательная среда вуза как совокупность условий реализации учебного процесса является инновационной, когда в ней обеспечивается полноценная поддержка самостоятельной работы обучающихся и профессиональной и научной деятельности преподавателей. При формировании инновационной образовательной среды выбор конкретных сетевых ресурсов должен опираться на единый междисциплинарный подход, позволяющий учитывать особенности коммуникации в разных средах, с различными участниками и задачами. Кроме того, при выборе этих ресурсов необходимо учитывать особенности использования и создания специальных текстов, которые являются основным источником извлечения информации и формирования знаний [8].

Таким образом, процесс использования сетевых ресурсов при подготовке филолога предполагает:

- 1) определение комплекса лингвистических ресурсов, использование которых необходимо для конкретного специалиста – исследователя, преподавателя, студента;
- 2) установление методов и приемов исследования и обучения на основе анализа всего арсенала средств информационных технологий и лингвистических ресурсов;
- 3) решение различных исследовательских и учебных задач в рамках единой высокотехнологичной образовательной среды вуза.

Под лингвистическими ресурсами обычно понимаются те средства лингвистической поддержки и сопровождения, которые используются для представления информации об исследуемом естественном языке и обрабатываемых текстах (лексикографические базы данных, онтологии, тезаурусы и пр.) [ср. 9, с. 97–99], для представления знаний и данных в системах обработки информации, для решения задач извлечения эмпирической информации. Кроме того, к лингвистическим ресурсам относят и собственно языковые источники (полнотекстовые базы данных, национальные и специализированные корпуса текстов), представляющие собой источник знаний о языках. Национальные корпуса текстов являются важным сетевым лингвистическим ресурсом, поскольку их сбалансированность по типам и жанров текстов, по времени написания и объемам выборок обеспечивает возможность использования таких корпусов в качестве эталонов для оценки степени распространенности и/или терминологичности конкретных лексических единиц. Соответственно, национальные корпуса текстов могут использоваться для решения конкретных исследовательских и методических задач.

Перечисленные выше ресурсы необходимы как для обучения, так и для обработки информации на естественном языке, а также решения задач извлечения эмпирической информации. Особое место в этом комплексе ресурсов занимают ресурсы лексикографические, предоставляющие возможность оперативного извлечения информации из различных источников: терминологических баз и банков данных, электронных словарей и словарей автоматизированных систем обработки текстов, предметно-ориентированных корпусов текстов и т.п.

При подготовке филолога именно лексикографические ресурсы являются базой поддержки обучения и формирования языковых компетенций, поэтому дальше подробно будут рассматриваться именно они. Подобные ресурсы можно разделить на те, в которых значения единиц описываются с помощью дефиниций, и те, в которых значение конкретной единицы описывается как ее место в системе семантических отношений между единицами. К первым относятся активно разрабатываемые терминологические базы/банки данных (ТБД). Эти базы представляют собой автоматизированные хранилища, в которых термины снабжаются дополнительной информацией как лингвистического (сочетаемость, частотность, синтаксические и семантические характеристики), так и экстралингвистического (нормативность, стандартизованность и т.п.) типа. «В зависимости от цели создания ТБД их можно разделить на две группы: ориентированные на обеспечение работ по переводу научно-технической литературы и документации и предназначенные для обеспечения информацией о стандартизованной и рекомендованной терминологии» [10, с. 284].

В свою очередь терминологические ресурсы по степени универсальности и доступности можно разделить на государственные (например, поддерживаемые Комиссией ЕС или конкретными государственными проектами) и инициативные, разрабатываемые корпорациями или исследовательскими группами.

Терминологический банк данных EuroTermBank [11] представляет собой один из самых мощных государственных терминологических банков, охватывая все языки Европейского союза, русский и китайский языки, такие «малые» языки Европы, как баскский, фарерский, галицийский, и латынь, всего 33 языка. В основную словарную базу к 2008 году было включено 1.240.000 словарных статей (5 миллионов терминов) и 325 000 аббревиатур и акронимов. Коды предметных областей, приписываемые каждой единице словаря, основаны на универсальной классификации Леноха.

При создании и ведении этого банка терминов ввод информации был организован в каждом из переводческих бюро ЕС в соответствии с его собственными правилами и подходами и в зависимости от различных соглашений об использовании и методов сотрудничества каждого языкового сообще-

ства и каждой страны. Разнообразие подходов, требований и стандартов определило целесообразность объединения всех отдельных баз терминов в согласованную базу, разрешающую постоянный ввод материала приблизительно 5000 переводчиков из учреждений ЕС. При таком подходе пополнение словарной базы данных происходило за счет работы терминологического бюро (в Брюсселе и Люксембурге); предложения, поступающие от переводчиков, систематизировались группой Eurodicautom, кроме того, часть информации поступала по контрактам от частных компаний и экспертов в отдельных областях знаний. Обновление системы происходило еженедельно [12].

В 2008 году в Европейском парламенте было принято решение создать особую структуру, предназначенную для координации исследований в области терминологии, согласования данных, получаемых из различных источников. Кроме того, для сохранения этих данных предполагалось использовать формат IATE (InterActive Terminology for Europe) – специализированного лингвистического ресурса, реализованного в виде терминологической реляционной базы данных. Кроме того, целью этой новой структуры была поддержка сотрудничества с учреждениями ЕС при создании и ведении новой базы данных, в которую вошли миллионы терминов, извлеченных из других баз и импортируемых без всякого анализа или предварительной фильтрации. Ведение базы предусматривало удаление устаревших и вышедших из употребления терминов, а также дубликатов недавно добавленных единиц. В качестве такой структуры Европейским парламентом организован отдел по координации терминологии TermCoord, который и осуществляет доступ к терминологии ЕС через общедоступный сайт и бесплатные инструментальные средства, а также через Межведомственный терминологический портал EurTerm [12].

Для поддержания базы данных в актуальном состоянии в описываемом ресурсе используются различные инструментальные средства: процедуры извлечения терминов из текста, макросы для облегчения предварительного хранения терминов в процессе перевода, серверы поиска для визуального просмотра сотен связей со специализированными импортированными глоссариями, веб-страницы, концентрирующие связи со всеми справочниками всех организаций ЕС. Некоторые из этих ресурсов (DocHound и GlossaryLinks) были опубликованы на сайте отдела и используются внешними подрядчиками, переводящими до 30 % текстов Европарламента.

Планируется, что терминологический портал EurTerm, разработанный отделом TermCoord, будет иметь доступ к внутреннему формату IATE, централизованный доступ ко всем межведомственным и мировым терминологическим ресурсам и банкам данных, а также доступ к платформам коммуникации на уровне конкретных языков (терминологии wikis и форумы). Кроме того, он будет обеспечивать доступ к инструменту QUEST для поиска терми-

нологии внутри терминологических хранилищ ЕС, этот инструмент осуществляет поиск в наиболее важных национальных терминологических базах данных в дополнение к IATE, переводческой памяти Euramis и EurLex [12].

В результате проведенных исследований и организационных мероприятий в этом лингвистическом ресурсе было объединено 133 локальных ресурса, разработанных в различных бюро перевода. Предусмотрены связи с 4 базами терминологических данных (базой IATE, базой данных Латвийской академии наук TermNet.lv, открытым терминологическим словарем сельскохозяйственной академии г. Щецин OSTEN, венгерской базой MoBiDic, поддерживающей переводы с венгерского языка). В ресурсе 2 650 976 терминов (число постоянно увеличивается), 710 705 словарных статей, 221 512 дефиниций на 33 языках [13]. Экспорт данных IATE доступен как отдельный дамп-файл для скачивания на соответствующем сайте Web или на открытом портале данных (Open Data Portal), начиная с февраля 2015 г. через инструмент IATEExtract, который разрешает выбор конкретных языков. Данные в словарной статье структурируются на трех уровнях: (i) уровень понятий, независимый от языка, (ii) уровень языка со специфической информацией для каждого языка и (iii) уровень терминов.

Структура информации в базе данных EuroTermBank предполагает различные опции выбора исходного языка и языка перевода, предметной области, формы представления информации. При выборе конкретных опций предоставляется информация о вариантах перевода в различных предметных областях и о зафиксированных в базе данных словосочетаниях. На рис. 1 представлена структура выдачи информации для перевода с русского языка на английский. Обращение к словарю осуществляется бесплатно.

Использование ресурса в его современном виде позволяет:

- осуществлять поиск терминов в различных источниках;
- выявлять в собственных текстах те слова и словосочетания, которые могут быть терминами (в современной терминологии – кандидатами в термины), и автоматически их извлекать;
- просматривать варианты перевода слова и словосочетания в разных предметных областях;
- искать термины в нескольких языках одновременно;
- уточнять переводы и делиться информацией с другими пользователями.

Доступ к ресурсу может осуществляться непосредственно из Microsoft Word.

В свою очередь система IATE (InterActive Terminology for Europe), реализованная как реляционная база данных и связанная с ресурсом EuroTermBank, жестко ориентирована на языки Евросоюза и поддерживает только их.

Информация о переводе по запросу пользователя включает возможность выбора пары языков из 25 возможных, выбора конкретной предметной области и подобласти в ней в соответствии с принятой в базе классификацией, определение типа лексической единицы – термин или аббревиатура.

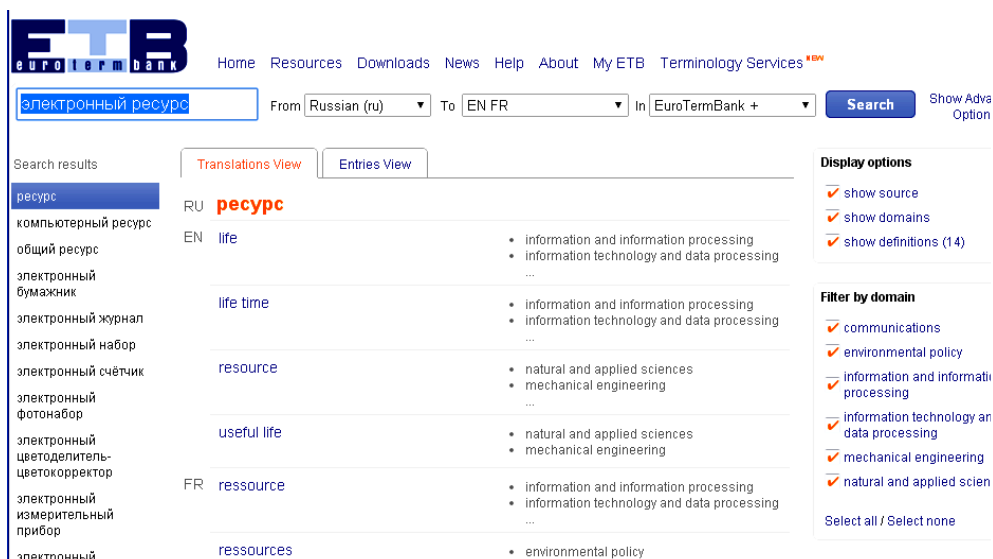


Рис. 1. Структура выдачи информации о переводе в базе данных EuroTermBank (фрагмент)

Ресурс базы терминов EuroTermBank, подробно описанный выше, может рассматриваться как опробованная модель многоязычного сетевого ресурса, создание которого актуально как для языков национальных республик России, так и для языков Таможенного союза, поскольку может обеспечить корректную терминологическую и лексикографическую поддержку для перевода документов в различных областях сотрудничества. Целесообразность разработки подобного ресурса в области формирования ресурсов образовательной среды также несомненна.

В то же время следует иметь в виду, что лексикографические ресурсы, подобные описанным выше, включают главным образом терминологию, привлекаемую в результате стандартизации, и (несмотря на огромные объемы) не способны охватить всю терминологию, особенно для активно развивающихся областей знаний. Сегодня основными недостатками любых терминологических ресурсов являются: высокая стоимость их создания и длительное время, необходимое для его разработки: недостаточный охват терминологии, особенно для оперативной фиксации и корректного номинирования самых современных понятий; недостаточность совместного использования терми-

нологических ресурсов и отсутствие механизмов для привлечения терминологов-практиков.

Новым «облачным» ресурсом, предназначенным для устранения этих недостатков, является разработка сервисной модели TaaS (Terminology as a Service), задачей которой является автоматизация основных этапов терминологической работы и оперативность создания и ведения терминологических баз данных [14]. Модель TaaS основана на принципе взаимности: каждый из тех, кто пользуется этим сервисом, вносит свою лепту в формируемые словарные системы. Пользователи обрабатывают документы, используя этот сервис, уточняют и обогащают получаемые терминологические данные, которые затем могут совместно использоваться и передаваться другим пользователям, а также вноситься в накапливаемые банки терминов.

Соответственно, TaaS предлагает набор функционально совместимых «облачных» сервисов, объединенных в последовательности выполнения работ:

- автоматическая идентификация кандидатов в термины в предоставляемых пользователем одноязычных документах;
- поиск эквивалентов перевода для извлеченных одноязычных кандидатов в термины;
- извлечение кандидатов в переводной эквивалент из параллельных или сопоставимых данных Интернета для терминов, отсутствующих в известных ресурсах, с использованием методов извлечения одноязычных терминов и выравнивания пар кандидатов в термины.

При этом переводные эквиваленты извлекаются:

- из терминологических банков в режиме онлайн;
- автоматически обрабатываемой многоязычной терминологии, извлекаемой из сопоставимых и параллельных ресурсов Интернета;
- коллекций терминов, созданных пользователями платформы.

Терминологические и, шире, лингвистические ресурсы используются как специалистами в области языка и перевода, так и различными программами автоматической обработки текстов, поэтому в платформу TaaS включено средство поиска терминов API для доступа подобных систем к терминологическим сервисам и данным [14]. Этот проект развивается очень активно и выступает перспективным лексикографическим ресурсом, ведение которого является коллективной терминологической работой всех свободно регистрируемых участников.

Метод реализации проектов, осуществляемых с привлечением различных заинтересованных участников, не обязательно специалистов в области лексикографии, терминоведения или перевода, принято называть краудсорсингом (*crowdsourcing*). Под краудсорсингом понимается процесс, который

вовлекает в исследование и сбор информации группу людей (часто называемую толпой – *crowd*), что способствует достижению цели конкретного проекта, поскольку нагрузка распределяется между членами группы [15]. Одним из самых ранних примеров активного участия широкой публики в создании словаря был Оксфордский словарь английского языка (*Oxford English Dictionary – OED*), при создании которого редакционная коллегия OED просила всех посылать свои дополнения, содержащие слова и примеры использования [16].

Разработка платформы TaaS подтверждает, что краудсорсинг может успешно использоваться в лексикографии – не как заключительный или главный этап создания словаря, но как метод фильтрации и обработки данных перед тем, как лексикографы и терминологию решают вопрос об их введении в реальный словарь [17].

В то же время существуют лингвистические ресурсы, для создания которых привлекаются эксперты самого высокого уровня, развитие таких ресурсов происходит в течение многих лет и накопленная в них информация является актуальной для филолога. Примером такого ресурса является разработка Принстонского университета, начатая еще в 1984 году – *WordNet*, сетевая база данных, представляющая собой очень важный лингвистический ресурс для филолога, по сути – компьютерный тезаурус, размещенный в сети Интернет [18]. В этой базе данных слова английского языка (существительные, глаголы, прилагательные и наречия) сгруппированы в наборы когнитивных синонимов (*синсеты*), каждый синсет номинирует отдельное понятие и является узлом семантической сети. Связи между синсетами определяются концептуальными, семантическими и лексическими отношениями. Поиск информации о семантически связанных словах и понятиях поддерживается специальным браузером. Система включает 117 000 синсетов (наборов синонимов), каждый из которых связан с другими небольшим количеством концептуальных отношений. Кроме того, каждый синсет содержит краткую дефиницию (*глоссу*) и не менее одного короткого предложения, иллюстрирующего использование элементов синсета. Многозначные словоформы представлены в отдельных синсетах, количество которых соответствует числу разных значений. Таким образом, каждая пара типа форма – значение является уникальной.

Наиболее часто используемым отношением между синсетами является иерархическое отношение (*гиперонимия, гипонимия* или отношение «является экземпляром»). Этот тип отношений связывает более общие синсеты с более частными. Все иерархии синсетов, включающих существительные, в конечном счете доходят до корневого узла (*сущности*). Гипонимическое отношение является транзитивным. Вторым типом концептуальных отношений между узлами-синсетами является *меронимия*, отношение типа часть–целое. Части наследуют свойства целого, но не наоборот.

WordNet Search - 3.1
 - [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
 Display options for sense: {frequency} {offset} <lexical filename > [lexical file number] (gloss)
 "an example sentence"
 Display options for word: word#sense number (sense key)

Noun

- (12){13352793} <noun.possession>[21] [S:](#) (n) **resource#1 (resource%1:21:00::)**
 (available source of wealth; a new or reserve supply that can be drawn upon when needed)
- (8){05161923} <noun.attribute>[07] [S:](#) (n) **resource#2 (resource%1:07:00::)** (a source of aid or support that may be drawn upon when needed) "*the local library is a valuable resource*"
- (2){05641705} <noun.cognition>[09] [S:](#) (n) **resource#3 (resource%1:09:00::), resourcefulness#2 (resourcefulness%1:09:00::), imagination#3 (imagination%1:09:03::)** (the ability to deal resourcefully with unusual problems) "*a man of resource*"

Рис. 2. Структура выдачи полной информации о существительном в системе WordNet

Иерархические отношения между синсетам глаголов устроены иначе: глаголы, синсеты которых расположены ближе к основанию деревьев (тропонимы), выражают более специфические способы, характеризующие событие. Конкретный способ зависит от параметров семантического поля; сюда входит объем, скорость или интенсивность. Глаголы, описывающие события, которые обязательно и однонаправленно предполагают друг друга, связаны между собой.

Основным отношением, фиксируемым для прилагательных, является антонимия, пары «непосредственных» антонимов отражают сильную семантическую связь между ними. У относительных прилагательных указываются существительные, для которых они являются производными. В WordNet включено только небольшое число частотных наречий, поскольку большинство наречий в английском языке выводятся из прилагательных путем морфологической аффиксации.

Особым образом в этой сети фиксируются отношения между словами разных частей речи. В большинстве случаев отношения устанавливаются между словами, принадлежащими одной и той же части речи. Соответственно, в сети есть 4 отдельные иерархии, между узлами которых заданы связи:

морфосемантические для однокоренных слов, принадлежащих разным частям речи, отношения, соответствующие семантическим ролям.

Более подробное описание в терминологии семантических ролей дает другой сетевой ресурс – FrameNet, разрабатываемый в Международном институте информатики в Беркли, Калифорния, теоретической основой описания является фреймовая семантика, идея которой разработана Ч. Филлмором [19], он был лидером проекта с его начала в 1997 году. Семантический фрейм представлен в этом проекте как концептуальная структура, особым формализованным образом описывающая события, отношения, объекты и участников. В базе данных FrameNet содержится около 1 200 семантических фреймов, 13 000 лексических единиц (под лексической единицей понимается пара типа слово – дефиниция; многозначные слова представлены несколькими парами) и более 190 000 предложений, формирующих базу примеров [20].

Эта лексическая база может использоваться как людьми, так и системами обработки информации. С точки зрения поддержки обучения она представляет собой словарь, содержащий более 10 000 значений слов, большая часть сопровождается аннотируемыми примерами, которые демонстрируют значение и употребление. Для исследований в области обработки текстов на естественных языках есть возможность использовать в качестве эталонного массива более 170 000 предложений, которые аннотированы вручную. Этот массив обеспечивает уникальный набор данных для маркировки семантической роли, используемой в приложениях типа информационного поиска, машинного перевода, опознавания событий, смысловому анализу и т.д.

Группа разработчиков системы FrameNet установила более 1000 семантических фреймов и установила их связи, что позволяет связывать более общие и более конкретные фреймы и обеспечивает базу для умозаключений о событиях и преднамеренных действиях. Поскольку фреймы, по сути, являются семантическими, они подобны в различных языках, и аналогичные системы разрабатываются сейчас для испанского, немецкого, китайского и японского языков [20].

Таким образом, можно утверждать, что сетевые лингвистические ресурсы представляют собой важный исследовательский, образовательный и методический ресурс, использование которого должно быть неотъемлемым условием подготовки филолога.

В принципе, все ресурсы высокотехнологичной образовательной среды можно разделить на информационные, технологические (программные) и технические. Такое деление ресурсов на виды не зависит ни от типа пользователей, ни от их принадлежности к гуманитарной или естественно-научной сфере образования и науки, не зависит и от конкретных задач, решаемых в рамках образовательных программ или исследовательских проектов. Предметная ориентированность информационных ресурсов определяет целесооб-

разность их иерархизации, т.е. выделения терминальных (базовых) узлов и узлов высокого уровня (универсальных). При этом терминальные узлы (терминологические базы или корпуса текстов) должны соответствовать реализуемым в вузе специализациям образования, т.е. достаточно узким областям знаний и релевантным именно для конкретной образовательной среды. При этом «узость» должна определяться не связью с конкретным факультетом или институтом, а учебной и научной спецификой.

Иерархическая структура информационных и технологических ресурсов предполагает, что на верхних уровнях (уровнях общего доступа) должны быть максимально универсальные ресурсы, использование которых не связано со специализацией исследования и обучения. Эта же иерархическая структура должна реализоваться для каждого конкретного (терминального) ресурса. Так, например, при единой системе машинного перевода, ориентированной на задачи обучения и перевода, выделяются конкретные системы автоматических словарей, соответствующие областям знаний. В предметной области «Филология» общим словарем верхнего уровня является словарь общефилологических терминов; на следующем уровне выделяются подобласти лингвистики, литературоведения и образовательных технологий в филологии, для каждой из которых должен формироваться общий словарь и система предметно-ориентированных словарей (фонетика, лексикология, лексикография, морфология и т.д. для предметной области «Лингвистика»). Эту систему, которая требует совместной поддержки, для решения своих задач смогут использовать специалисты и студенты разных факультетов [21, 22].

Список литературы

1. Климзо Б.Н. Ремесло технического переводчика. Об английском языке, переводе и переводчиках научно-технической литературы. – 2-е изд., перераб. и доп. – М.: Р. Валент, 2006.
2. Кривых Л.Д., Рябичкина Г.В., Смирнова О.Б. Технический перевод. – М.: Форум - Инфра-М, 2008.
3. Сальмон Л. Теория перевода. История, наука, профессия. – СПб.; Астана, 2007.
4. Alwert K., Hoffmann I. Knowledge Management Tools / K. Mertins, P. Heisig, H. Vorbeck (eds.) // Knowledge Management. Concepts and Best Practices. – Berlin; Heidelberg; New York, 2003. – P. 114–150.
5. Логичев С.В. Каталог лингвистических программ и ресурсов в Сети [электронный ресурс] – URL: <http://rvb.ru/soft/catalogue/catalogue.html> (дата обращения: 2.10.2015).

6. Усталов Д.А. Каталоги лингвистических ресурсов: состояние и перспективы [Электронный ресурс] – URL: <http://www.moluch.ru/archive/47/5955/> (дата обращения: 2.10.2015).

7. Загорюлько Ю.А., Боровикова О.И., Кононенко И.С. Обеспечение содержательного многоязычного доступа к лингвистическим информационным ресурсам на основе технологии порталов знаний // Известия Томского политехнического университета. – 2011. – Т. 318, № 5 – С. 99–104.

8. Беляева Л.Н. Лингвистические ресурсы информационной образовательной среды: состав, структура, функции // Известия РГПУ им. А.И. Герцена. – 2014. – № 171. – С. 47–52.

9. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика / Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ, А.А. Носков, О.В. Пескова, Е.В. Ягунова. – М.: МИЭМ, 2011.

10. Лейчик В.М. Прикладное терминоведение и его направления // Прикладное языкознание: учеб. / под ред. А.С. Герда. – СПб.: Изд-во С.-Петерб. ун-та, 1996. – С. 276–286.

11. Towards Consolidation of European Terminology Resources. Experience Recommendations from EuroTermBank Project / Ed. by: S. Rirdance, A. Vasiljevs. – Riga: Tilde, 2006. – 123 p.

12. Maslias R. Combining EU Terminology with Communication and Ontology Research // Terminology and Knowledge Engineering 2014, 19–21 June 2014. – P. 48–56.

13. Eurotermbank. – URL: http://www.eurotermbank.com/Collection_list.aspx?langu=bg (дата обращения 6.10.2015).

14. Vasiljevs A., Pinnis M., Gornostay T. Service model for semi-automatic generation of multilingual terminology resources // Terminology and Knowledge Engineering: Proceedings of the 11th Conference on Terminology and Knowledge Engineering, Germany, 19–21 Jun 2014. – Berlin, 2014. – P. 67–76.

15. Howe J. Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business. – New York: Crown Publishing Group, 2008.

16. Lanxon N. How the Oxford English Dictionary started out like Wikipedia. – URL: <http://www.wired.co.uk/news/archive/2011-01/13/the-oxford-english-wiktionary> (дата обращения: 6.10.2015).

17. Čibej J., Fišer D., Kosem I. The role of crowdsourcing in lexicography // Proc. of the fourth biennial conference on electronic lexicography, eLex 2015: Linking lexical data in the digital age. – Sussex, United Kingdom, 11–13 August 2015. – P. 72–79.

18. WordNet. A lexical database for English. – URL: <http://wordnet.princeton.edu/> (дата обращения: 6.10.2015).

19. Atkins S., Fillmore C. Starting where the dictionaries stop: The challenge for computational lexicography / B.T.S. Atkins, A. Zampolli (Eds.) // Com-

putational Approaches to the Lexicon. – Oxford: Oxford University Press, 1994. – P. 349–393.

20. FrameNet. – URL: <https://framenet.icsi.berkeley.edu/fndrupal/about> (дата обращения: 6.10.2015).

21. Беляева Л.Н. Лингвистические ресурсы информационной образовательной среды: состав, структура, функции // Известия РГПУ им. А.И. Герцена. – 2014. – № 171. – С. 47–52.

22. Беляева Л.Н. Лингвистическое обеспечение высокотехнологичной образовательной среды: состав и структура // Коммуникация в поликодовом пространстве: лингво-культурологические, дидактические, ценностные аспекты: материалы междунар. науч. конф. – СПб.: С.-Петербург. политехн. ун-т Петра Великого, 2015. – С. 4–6.

Получено 11.11.2015

L.N. Beliaeva

WEB-BASED LINGUISTIC RESOURCES FOR A PHILOLOGIST

The paper considers research and educational potential of web-based linguistic resources which are necessary for information mining and data extraction. The recommended resources are presented.

Key words: *web-based linguistic resources, termbases, information mining, term extraction and translation.*