УДК 81'322

К ВОПРОСУ ОБ ЭЛЕКТРОННЫХ РЕСУРСАХ ПРОФЕССИОНАЛЬНОЙ ЛЕКСИКИ

С.О. Шереметьева, П.Г. Осминин, Е.С. Щербаков

Описывается методика построения базы знаний для анализирующего модуля системы автоматизированного реферирования научно-технических текстов. База знаний строится на основе анализа предметной области и ориентирована на автоматизацию определения содержания реферата по тексту статьи.

Ключевые слова: автоматизированное реферирование, определение содержания, база знаний, научно-технический текст.

1. Введение

В настоящее время, когда темпы глобализации требуют оперативного обмена научно-технической информацией в международном масштабе, проблема перевода профессиональных текстов, качество которого во многом определяется корректностью перевода используемых лексических единиц, стоит особенно остро. При этом стремительный технический прогресс приводит как к постоянному появлению новых терминов, так и изменениям в использовании общеупотребительной лексики. Переводчик, как правило, не обладающий экспертными знаниями в предметной области переводимого текста, затрачивает на перевод терминов около 75 % времени, а процент ошибок как при переводе терминологии, иногда, и общеупотребительной лексики в профессиональном контексте достигает 40 % [1].

Бумажные словари – до недавнего времени основной ресурс переводчика, громоздки, требуют больших временных затрат при поиске переводных эквивалентов, как правило, страдают низкой покрываемостью и не отражают актуальных изменений инвентаря профессиональной языковой коммуникации.

С внедрением информационных интернет-технологий, электронных словарей, систем машинного перевода (МП) и других электронных ресурсов переводчик получил возможность сократить трудоемкость и время выполнения переводов, повысить качество своей продукции. При этом на первый план выдвигаются как проблемы повышения качества электронных ресурсов, особенно, словарей профессиональной лексики, так и проблемы корректного применения пользователями уже имеющихся и ставших широкодоступными интернет-ресурсов.

2. Анализ современных переводческих интернет-ресурсов

Пользователю важно уметь ориентироваться в огромном количестве переводческих ресурсов в интернете (например, поисковая система Google на запрос «Словарь» дает около 8 млн результатов), осознавать степень их надежности и уметь корректно ими пользоваться. На рис. 1 приведена классификация интернет-ресурсов, которые могут оказаться полезными как для переводчиков или специалистов-ученых, так и для лексикографов, создающих словари для специализированных предметных областей. Существуют оф- и онлайновые электронные переводческие ресурсы. При этом офлайновые электронные ресурсы можно купить или скачать из интернета и пользоваться ими без соединения с интернетом, что можно считать их преимуществом. Однако обновление таких ресурсов, например словарей, бывает достаточно проблемным. Онлайновые электронные ресурсы существуют в сети интернет и обновляются разработчиками, обеспечивая пользователям доступ к текущим версиям таких ресурсов.

Вспомогательный инструментарий, как правило, предназначен для лексикографов, а также разработчиков систем электронных словарей и

Рис. 1. Классификация электронных переводческих ресурсов

автоматического перевода и используется, в частности, для автоматизации отбора вокабуляров одно- и многоязычных лексиконов. Это такие, например, программы как вычислители n-грам, сортировщики, конкордансы, экстракторы лексических единиц, межъязыковые выравниватели и т. д.

Электронные словари и, как их разновидность, *справочники* можно разделить на два основных вида: электронные копии бумажных лексиконов и электронные словари с пользовательским интерфейсом, основанные на базе данных или знаний разной глубины.

Электронные копии бумажных лексиконов и других справочников существуют в форматах .pdf, .djvu и др. [2] и снабжены сервисом поиска. Такие словари отличаются от бумажных только отсутствием физического веса, интернет-доступностью и сокращением времени поиска эквивалентов.

Электронные словари с пользовательским интерфейсом, основанные на базе данных или знаний разной глубины, могут кроме собственно перевода лексической единицы давать морфологическую информацию, например, множественное число или всю парадигму словоформ, а также перевод лексемы в разных контекстах и учитывать предметную область (математика, механика, экономика и т. д.). Такие словари могут выдавать эквиваленты на нескольких языках, что, несомненно, является их достоинством.

Проблемой таких словарей, особенно в специализированных областях, остается покрываемость, а иногда и корректность. Заявленные разработчиками возможности словарей довольно часто не соответствуют действительности.

Очень популярны словари открытого типа, например, Abbyy Lingvo [4], MultiTran [5], куда любой пользователь может внести переводные эквиваленты. Преимуществом открытых словарей является все время пополняемый обширный вокабуляр, что обеспечивает лучшую покрываемость словаря, а одним из недостатков – то, что по запросу пользователя часто выдается слишком много вариантов перевода лексемы и (или) ее компонентов, при этом наличие многочисленных контекстов не всегда помогает. Классификация терминов по предметным областям не всегда содержит требуемый пользователем раздел и поэтому бывает достаточно трудно осуществлять навигацию по словарю и принимать решения относительно предложенных вариантов. Например, в Англо-русском научно-техническом словаре, который входит в состав электронного словаря Abbyy Lingvo, термин «*роевое представление частицы*» отсутствует, но даются следующие эквиваленты компонентов термина: *«рой» – «swarm», «представление» – «conception», «expression», «representation», «частица» – «bit», «fraction», «particle»*. В словаре Multitran предлагаются следующие варианты: *«рой» – «swarm», «представление» – «performance», «configuration», «частица» – «shard», «corpuscle»*. Ориентироваться в таком количестве вариантов пользователю трудно и составить правильный перевод многокомпонентного термина проблематично. Недостатком открытых словарей является и то, что предложенный вариант перевода не всегда может оказаться корректным, поскольку внесен не профессиональным переводчиком или лексикографом.

Системы «Память переводчика» (накопители), например, Across [7], представляют собой базу данных для хранения пар сегментов текста «оригинал – перевод». Сегментом может быть фраза, предложение, абзац. При переводе текста, текущий сегмент автоматически сравнивается с тем, что имеется в базе данных. Если в базе данных имеется совпадение, то переводчику предлагается переведенный сегмент. Переводчик может использовать предложенный перевод или отредактировать его и занести в базу данных. Если совпадения нет, то переводчик переводит текущий сегмента текста и он заносится в базу данных. Накопители эффективны при переводе однотипных текстов, содержащих много повторений, например, завещаний или инструкций по эксплуатации определенного типа оборудования. При малом количестве повторений даже в одной и той же предметной области накопители малоэффективны.

Что касается *систем машинного перевода*, то даже среди платных невозможно найти систему, обеспечивающую корректный перевод каждого типа профессиональных текстов. Качество же перевода специальных текстов с помощью бесплатных онлайновых систем МП страдает как от недостаточной корректности и покрываемости лексиконов, так и от проблем, связанных с разрешением неоднозначности значений лексических единиц и синтаксических структур. Например, системой машинного перевода PROMT [3] термин «*роевое представление частицы*» переводится *«royevy representation of a particle*» (термин «*роевое»* в лексиконе системы отсутствует и слово транслитерируется), а такой термин из области математического моделирования как «далекие младшие разряды» переводится как «far younger categories», что неправильно.

Недостаточная покрываемость и вызывающая сомнение корректность предлагаемого словарем или системой МП иноязычного эквивалента может быть компенсирована обращением к электронным онлайновым *корпусам* одноязычных, а также *параллельных и квазипараллельных текстов* релевантной предметной области. Параллельные тексты (корпусы текстов) – это тексты на разных языках одного и того же содержания. *Квазипараллельные тексты* имеют сходное, но не полностью совпадающее содержание, например, научные статьи на разных языках по одной теме, но не являющиеся переводами друг друга.

Одноязычные специальные тексты на исходном языке позволяют переводчику правильно оп-

ределить значение используемой единицы в контексте, чтобы затем найти ее корректный эквивалент в параллельном (квазипараллельном) корпусе текстов. Недостатками использования параллельных (квазипараллельных) текстов является то, что, во-первых, для каждой конкретной узкоспециальной области их трудно, а часто, и невозможно найти, и, во-вторых, процесс эффективного поиска переводных эквивалентов в параллельных (квазипараллельных) текстах далеко не тривиален и требует как знания иностранных языков, так и специальной методики. При этом для эффективного применения корпусного метода определения межъязыковых эквивалентов часто требуется вспомогательный инструментарий.

Таким образом, существующие переводческие интернет-ресурсы наряду с их несомненным преимуществом оперативности и доступности еще далеки от совершенства и требуют специальных навыков их использования.

3. Использование интернет-ресурсов для корректного перевода профессиональной лексики

В настоящее время переводчик или профессионал в любой другой области, не использующий компьютерный инструментарий для повышения оперативности своей деятельности, неконкурентоспособен на рынке труда. При этом человек, который остается необходимым звеном переводческого процесса, должен использовать интернет-ресурсы с известной долей осторожности, что особенно важно при переводе узкоспециальной лексики.

Как показывает анализ интернет-ресурсов, проведенный в предыдущей главе, при поиске межъязыковых терминологических эквивалентов:

- не рекомендуется пользоваться онлайновыми системами МП;
- следует помнить, что онлайновые словари открытого типа, такие как MultiTran, не гарантируют правильности перевода, поскольку создаются пользователями, не всегда являющимися профессиональными переводчиками или лексикографами;
- следует отдавать предпочтение терминологическим эквивалентам (в случае их наличия, что, к сожалению, часто оказывается проблематичным) в электронных копиях бумажных специализированных лексиконов, созданных профессиональными лексикографами;
- наиболее надежным источником современных корректных межъязыковых эквивалентов являются онлайновые *корпусы параллельных и квазипараллельных текстов*. При этом важно помнить, что иноязычные параллельные (квазипараллельные) тексты должны быть написаны носителями языка или напечатаны в журналах, где обеспечивается редакция переводов экспертами и носителями языка. В противном случае переводные тексты не гарантируют отсутствие ошибок;
- при наличии нескольких вариантов перевода всего термина или отдельных его компонентов следует убедиться, что:
  - значение переводимого термина понято правильно, путем обращения к контексту исходного документа и электронным одноязычным справочникам по специальности;
  - *квазипараллельный иноязычный текст (корпус*) действительно соответствует предметной области переводимого текста. Рекомендуется найти в интернете статьи носителей языка, указанные в списке литературы переводимой статьи;
- для проверки правильности и коррекции переводных эквивалентов следует использовать поисковые системы интернета на языке перевода;
- в случае наличия переводов только отдельных компонентов термина рекомендуется использовать их в качестве ключевых слов.

Для примера рассмотрим поиск в интернете английского эквивалента термина «*роевое представление частиц*», упомянутого в предыдущем разделе, комбинируя приведенные выше варианты перевода частей этого термина, полученных с помощью словарей Abbyy Lingvo и MultiTran.

В системе Академия Google поиск по запросу «swarm, representation, particle» дает во втором результате поиска фразу «swarm representation of the quantum particle». Таким образом, термин «*роевое представление частицы*» переводится *как «swarm representation of the(a) particle»*. Статья [8], в которой поисковая система выделила этот термин, представляет собой перевод с русского, опубликованный известным издательством Springer, у которого есть корректоры-носители языка, поэтому перевод можно считать правильным.

Другой способ определения английского эквивалента термина «*роевое представление частицы*» – начать с поиска русской статьи, содержащей этот термин. Google по данному термину дает сведения о статье «Моделирование квантовой динамики через классическое коллективное поведение». По названию эта статья находится в библиотеке eLIBRARY.RU, где, в свою очередь, приводится ссылка на переводную версию данной статьи (параллельный текст), выполненную издательством Springer и уже упомянутую выше, в которой имеется перевод нашего термина.

Напомним еще раз, что сравнение двух параллельных и, особенно, квазипараллельных корпусов текстов для нахождения переводных эквивалентов вручную может быть времязатратным, требует знания английского языка. Эту процедуру можно облегчить применением вспомогательного инструментария, пример использования которого описан в следующем разделе.

Проведенный анализ переводческих ресурсов свидетельствует, что проблема создания эффективных электронных профессиональных лексиконов по-прежнему остро стоит на повестке дня и целью следующего раздела является внести определенный вклад в ее решение.

4. Методика разработки электронных словарей узкоспециальной лексики

В настоящем разделе предлагается методика разработки электронных словарей для перевода узкоспециализированных текстов и описывается ее реализация на примере разработки электронного словаря для области математического моделирования, основные этапы которой приведены ниже.

На *первом этапе* определяется предметная область, цель и пользователи словаря. Словарь предназначен, в первую очередь, для русско-английского перевода текстов по математическому моделированию с возможностью его использования как человеком (переводчиком или математиком), так и в качестве лексикографического компонента систем автоматической обработки текста [6]. Отличительной чертой представляемого словаря является то, что он, в отличие от других электронных словарей, обеспечивает поддержку пользователя при составлении запросов и допускает обработку целого текста статьи с выдачей английских эквивалентов всей использованной лексики.

На *втором этапе* определяется модель знаний: типы, объем и формализм представления лингвистической информации в словарной статье.

Словарные статьи русских лексических единиц и их английских эквивалентов одинаковы по своей структуре и содержат несколько следующих зон.

ЗОНА 1 перечисляет все морфологические формы лексемы, в которых она встречается в текстах по математическому моделированию.

ЗОНА 2 содержит имя семантического класса лексемы.

ЗОНА 3 содержит имя части речи.

ЗОНА 4 имеется только у предикатной лексики и содержит семантические падежи предиката, их ранги и модель управления.

ЗОНА 5 приводит ранжированные по частоте линейные клише совместной реализации в тексте предиката и его семантических падежей. Например, линейное клише вида (2 3 × 1 4), где 1, 2, 3, 4 – это коды семантических падежей, а символ «х» показывает место предиката, может соответствовать такой фразе: (2: *в статье*) (3: *подробно*) *рассматривается* (1: *проблема применения уравнений в частных производных*) (4: *для решения этой задачи*).

Знания для зон 4 и 5 русских и английских словарных статей извлекаются на основе глубокого семантико-синтаксического анализа корпуса текстов, описание которого выходит за рамки настоящей статьи.

Третий этап состоит в определении источника и состава русского вокабуляра с лингвистической информацией в соответствии с моделью знаний.

В качестве исходного материала нашего словаря использован корпус статей по математическому моделированию, опубликованных в «Вестнике ЮУрГУ» в 2008–2013 гг. Первичный вокабуляр словаря, в который включены как термины, так и другие лексические единицы, определен с помощью обработки русского корпуса экстрактором LanA-Key [10] с последующей проверкой извлеченных списков лексикографом. Экстрактор позволяет автоматически извлекать из текстов списки различных типов лексики (именных и глагольных групп, прилагательных, наречий и т. д.) длиной до 4 компонентов.

Отличительной чертой нашей методики является пополнение вокабуляра терминами из сети интернет посредством использования терминов первичного списка в качестве ключевых слов, например, в поисковой системе Google. Новые термины отбираются из представленных на двух первых страницах результатов поиска. Так, например, для ключевого термина «псевдообращение» из первичного списка были найдены следующие расширения: «псевдообращение сопряженной системы», «псевдообращение матриц», «псевдообращение Мура-Пенроуза» и др.

На четвертом этапе определяются источники английских эквивалентов и собственно английские эквиваленты для каждой единицы русского вокабуляра с лингвистической информацией о переводных эквивалентах в соответствии с моделью знаний. В качестве источников переводных знаний использовались как бумажные словари, так и интернет-ресурсы, описанные в разделе 2, в частности были отобраны английские тексты, параллельные и квазипараллельные статьи из русского корпуса.

Поиск английских эквивалентов единиц русского вокабуляра осуществляется по методике, описанной в разделе 3, с использованием англоязычной версии экстрактора LanA-Key для извлечения лексических единиц определенного типа из параллельных и квазипараллельных текстов и последующим сравнением полученных русских и английских списков лексем.

Особое внимание уделяется устранению проблем пользователя с выбором переводного эквивалента из нескольких возможных. Это достигается прежде всего тем, что многокомпонентные лексические единицы, составляющие основную часть вокабуляра словаря, в основном однозначны. В случае наличия для русской единицы синонимичных английских эквивалентов в словарь заносится только один, наиболее частотный, что достаточно для перевода текстов. Часть многозначных лексем в данной предметной области используется только в одном из своих значений, которое и заносится в

словарь. Лексемы, сохраняющие многозначность в предметной области, заносятся в словарь (и выдаются пользователю) в возможно узком однозначном контексте. Например, многозначное слово *«решение» («solution», «decision»)* содержится в таких словарных статьях как *«принимать решение»* (*«make decision»*), *«находить решение»* (*«find solution»*) и др.

Пятый этап предполагает спецификацию и программную реализацию словаря. Электронная оболочка нашего ресурса создана на основе повторного использования и адаптации программы TransDict [9] и включает в себя модуль базы лингвистических знаний, отвечающей сформулированной выше модели, с интерфейсами конечного пользователя и лексикографа и DLL для внешних приложений. Интерфейсы снабжены эффективными сервисами поддержки деятельности пользователя и лексикографа. Блок-схема словаря дана на рис. 2.

Шестой этап – введение знаний, полученных в результате выполнения третьего и четвертого этапов в электронную словарную оболочку через интерфейс лексикографа. В момент подготовки настоящей статьи представленный электронный словарь содержит 30 000 словарных статей.

5. Заключение

В статье проанализирован переводческий потенциал современных интернет-ресурсов и дана методика их эффективного использования. Основной акцент сделан на проблеме корректности переводных эквивалентов профессиональной лексики, которую предложено решать, комбинируя существующие электронные словари, параллельные и квазипараллельные корпусы текстов, а также поисковые системы интернета с привлечением компьютерного инструментария автоматической обработки текстов.

Описана методика создания электронного словаря для перевода узкоспециализированных текстов. Разработаны эффективный формат и программная оболочка лексикона с интерфейсами пользователя и лексикографа, позволяющие ис-

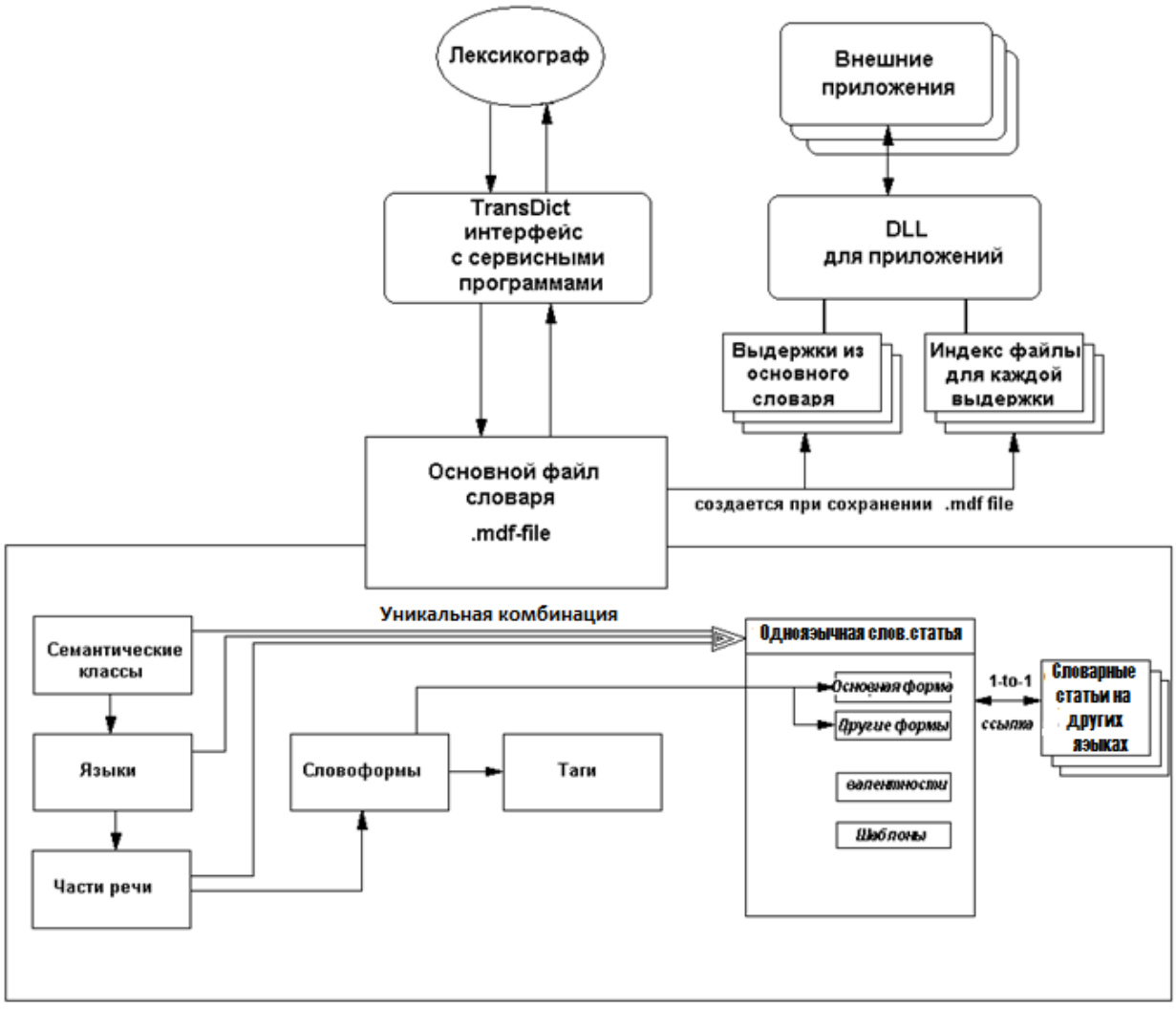

Рис. 2. Блок-схема электронного словаря

пользование ресурса как в качестве самостоятельного инструмента, так и в качестве лексикографического модуля систем автоматической обработки текста, например, машинного перевода. Действенность методики иллюстрируется на примере ее применения при создании электронного русско-английского лексикона предметной области математического моделирования.

Литература

1. Кудашев, И.С. Проектирование переводческих словарей специальной лексики / И.С. Кудашев. – Helsinki University Print, 2007. – 445 с.

2. Мюллер, В.К. Англо-русский словарь / В.К. Мюллер. – http://www.twirpx.com/file/486488/.

3. Онлайн-переводчик PROMT. – http://www.translate.ru.

4. Онлайн-словарь ABBYY Lingvo.Pro. – http://lingvopro.abbyyonline.com/ru.

5. Словарь Мультитран. – http://www.multitran.ru.

6. Шереметьева, С.О. К вопросу об автоматизации реферирования и аннотирования научно-технической литературы / С.О. Шереметьева // тр. междунар. науч.-практ. конф. «Лингвистика в контексте культуры». – Челябинск, 2012. – С. 304–309.

7. Across Personal Edition. – http://www.my-across.net/.

8. Ozhigov, Y.I. Simulating quantum dynamics in terms of classical collective behavior / Y.I. Ozhigov // Russian Microelectronics. – 2007. – V. 36, № 3. – P. 193–202.

9. Sheremetyeva, S. Application Adaptive Electronic Dictionary with Intelligent Interface / S. Sheremetyeva // Proceedings of the workshop on Enhancing and using electronic dictionaries in conjunction with the 20th International Conference on Computational Linguistics. COLING 2004. – Geneva, 2004. – P. 23–28.

10. Sheremetyeva, S. Automatic Extraction of Linguistic Resources in Multiple Languages / S. Sheremetyeva // Proceedings of NLPCS 2012, 9th International Workshop on Natural Language Processing and Cognitive Science in conjunction with ICEIS 2012. – Wroclaw, 2012. – P. 44–52.

Шереметьева Светлана Олеговна, доктор филологических наук, доцент, профессор кафедры «Лингвистика и межкультурная коммуникация», Южно-Уральский государственный университет (г. Челябинск). E-mail: linklana@yahoo.com.

Осминин Павел Григорьевич, аспирант кафедры «Лингвистика и межкультурная коммуникация» Южно-Уральский государственный университет (г. Челябинск). Научный руководитель – доктор филологических наук, профессор С.О. Шереметьева. E-mail: osperevod@gmail.com.

Щербаков Егор Станиславович, студент факультета «Лингвистика и межкультурная коммуникация», Южно-Уральский государственный университет (г. Челябинск). E-mail: dieggo@mail.ru.

Bulletin of the South Ural State University
Series "Linguistics"
2014, vol. 11, no. 1, pp. 57–63

ON ELECTRONIC RESOURCES FOR PROFESSIONAL LEXICON

S.O. Sheremetyeva, *South Ural State University, Chelyabinsk, Russian Federation, linklana@yahoo.com.*
P.G. Osminin, *South Ural State University, Chelyabinsk, Russian Federation, osperevod@gmail.com.*
E.S. Scherbakov, *South Ural State University, Chelyabinsk, Russian Federation, dieggo@mail.ru.*

The paper analyzes translation potential of current electronic resources and suggests an efficient way to use Internet for solving the problem of cross-linguistic equivalents for the lexical inventory of professional texts. The authors present a methodology for the development and the format of a multilingual and multipurpose electronic professional lexicon. The approach is illustrated on the example of electronic Russian-English lexicon for mathematical modeling.

Keywords: e-resources, internet, translation, professional lexicon.

References

1. Kudashev I.S. *Proektirovanie perevodcheskikh slovarey spetsial'noy leksiki* [Designing LSP Dictionaries for Translators]. Helsinki university print, 2007, 445 p.
2. Miuller V.K. *English-Russian Dictionary.* Available at: http://www.twirpx.com/file/486488/
3. *Onlain-dictionary ABBYY Lingvo.Pro.* Available at: http://lingvopro.abbyyonline.com/ru
4. *Dictionary Multitran.* Available at: http://www.multitran.ru
5. *Across Personal Edition.* Available at: http://www.my-across.net/
6. *Onlain-perevodchik PROMT.* Available at: http://www.translate.ru
7. Ozhigov Y.I. Simulating quantum dynamics in terms of classical collective behavior. *Russian Microelectronics*, 2007, V. 36, no. 3, pp. 193–202.
8. Sheremetyeva S. K voprosu ob avtomatizatcii referirovaniia i annotirovaniia nauchno-tekhnicheskoi literatury. *Trudy mezhdunarodnoi nauchno-prakticheskoi konferentcii «Lingvistika v kontekste kultury»*, Cheliabinsk, 2012, pp. 304–309.
9. Sheremetyeva S. Automatic Extraction of Linguistic Resources in Multiple Languages. *Proceedings of NLPCS 2012, 9th International Workshop on Natural Language Processing and Cognitive Science in conjunction with ICEIS 2012*, Wroclaw, Poland, 2012, pp. 44–52.
10. Sheremetyeva S. Application Adaptive Electronic Dictionary with Intelligent Interface. *Proceedings of the workshop on Enhancing and using electronic dictionaries in conjunction with the 20th International Conference on Computational Linguistics. COLING 2004*, Geneva, Switzerland, August, pp. 23–28.

Svetlana O. Sheremetyeva, PhD (Habilitation), professor of the Linguistics and Intercultural Communication department, South Ural State University (Chelyabinsk). E-mail: linklana@yahoo.com

Pavel G. Osminin postgraduate student of the Linguistics and Intercultural Communication department South Ural State University (Chelyabinsk). E-mail: osperevod@gmail.com

Egor S. Scherbakov, undergraduate student of the Linguistics faculty, South Ural State University (Chelyabinsk). E-mail: dieggo@mail.ru

Поступила в редакцию 24 января 2013 г.