Корпоративная переводческая сеть с использованием специальных электронных библиотек

© В.Е. Абрамов
ЗАО СКБ «ТЭЛКА»
abramval@yandex.ru

Н.Н. Абрамова, А.А Карнацкая, В.М. Рожков
НИЦИ при МИД России
nabramova@mid.ru, akarnatskaya@mid.ru, vrozhkov@mid.ru

Аннотация

В статье описывается создание корпоративной переводческой сети, объединяющей несколько автоматизированных рабочих мест (АРМ) переводчика, на платформе IBM Lotus Domino/Notes. Разработка этой сети вызвана необходимостью автоматизировать труд переводчиков, что поможет соответствовать современным требованиям качества и скорости перевода. Система построена на принципах памяти переводчика, когда используется большой корпус текстов на четырех языках (русском, английском, французском и испанском), переведенных вручную. Описаны компоненты АРМ переводчика, приводятся примеры работы системы.

1 Введение

В настоящее время достигнуты определенные успехи в области машинного перевода. Однако профессиональные переводчики практически не используют системы машинного перевода, мотивируя это неудовлетворительным качеством перевода. На постредактирование переведенного материала иногда можно затратить больше времени, чем на перевод по старинке без помощи программ. Как известно, ни одна из ныне существующих в мире систем перевода не может обеспечить уровень перевода, сравнимый с уровнем человека-переводчика.

В то же время, современному переводчику необходимы средства автоматизации, облегчающие его труд. Такие средства, называемые «память переводчика», «накопители переводчика» или «накопители переводов» стали создаваться начиная с 80-х годов прошлого века. Этому способствовали большие объемы накопленных к тому времени параллельных текстов на разных иностранных языках, переведенных вручную, а также прогресс в области вычислительной техники, позволивший создать информационные системы для накопления, хранения и поиска информации.

Появилось направление машинного перевода, основанное на принципе памяти переводчика. За рубежом основоположником этого направления явился японский профессор М. Нагао [6], а в России идеологом стал профессор Белоногов Г.Г., под руководством которого была создана система фразеологического перевода Retrans [3].

На аналогичных принципах строятся системы автоматизированного перевода, выполняющие в отличие от систем машинного перевода не полный перевод текста, а его фрагменты без формирования связного текста, оставляя за человеком значительную часть по переводу, согласованию и редактированию текста. На сегодняшний день известно несколько часто используемых систем автоматизированного перевода, например, Trados [7], OmegaT [12], SDLX [9], Wordfisher [14], Metatexis [10], DejaVu [8], Transit [11], TermStar [13]. В обзорной статье [5] и докладе [4] можно познакомиться с общими характеристиками и возможностями, а также особенностями их архитектуры и принципами работы.

Помимо программных систем помощь переводчику оказывают автоматические словари, среди которых автор обзора [5] отмечает Translatelt, PROMT VER-Dict , ABBYY Lingvo, Мультитран, Контекст.

2 Необходимость создания корпоративной переводческой сети

Однако, несмотря на некоторые успехи в области автоматизированного перевода, в реальной жизни переводчики с большим трудом могут воспользоваться этими разработками. Многие переводчики действуют по принципу «Omnia mea mecum porto» (все свое ношу с собой), стараясь иметь на своем компьютере (насколько позволяют технические возможности) как можно больше различных словарей, глоссариев, параллельных

Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.

текстов, программных продуктов, облегчающих процесс перевода.

В условиях глобализации современного мира выдвигаются более высокие требования к переводу. Постоянно происходит рост требующих перевода на иностранные языки материалов, так как расширяются международные связи, и как следствие, растет количество договорных актов и соглашений, заявлений в СМИ и т.п. К качеству перевода международных документов всегда предъявляются повышенные требования. Кроме того, переводчики ограничены во времени, поскольку переводные материалы должны появляться достаточно быстро после печати оригинала и даже в одно время с ним. На перевод накладываются довольно жесткие требования к используемой терминологии: термины, впервые появившиеся в основополагающих международных документах, таких как резолюции ООН, международные конвенции и договоры и переведенные на иностранные языки, в последующих документах должны переводиться таким же образом. То есть речь идет не столько о предоставлении переводчику переводных эквивалентов для ускорения процесса перевода, сколько о стандартизации этого процесса. Естественно, в организациях, имеющих штат профессиональных переводчиков, необходимо иметь корпоративную сеть, которая давала бы возможность всем переводчикам обращаться в единую информационную базу, чтобы каждый отдельный переводчик не тратил силы на создание собственной базы, а отдал бы свои лексические богатства, накопленные за долгие годы переводческой деятельности, для создания общей корпоративной базы данных.

3 Основные требования

Что же нужно переводчику для работы на современном уровне? Прежде всего персональный компьютер с достаточным объемом оперативной и дисковой памяти и высокой скоростью обработки информации, оснащенный DVD и периферийными устройствами (сканер, принтер, web-камера) и имеющий доступ в корпоративную сеть и сеть Интернет.

На рабочей станции (персональном компьютере) переводчика должен быть установлен хотя бы минимальный набор программ, позволяющих проводить обработку документов на русском и иностранных языках, включающий текстовый редактор, систему оптического распознавания текстов, электронные переводчики, клиент-серверное программное обеспечение (ПО) для совместной работы. Дополнительно на рабочую станцию можно установить системы автоматического реферирования текстов [1] и системы распознавания голоса. Базы данных с электронными словарями и параллельными текстами на разных языках должны быть в общем пользовании и располагаться на сервере.

4 Состав и принципы работы корпоративной переводческой сети

4.1 Особенности платформы IBM Lotus Domino/Notes

В качестве среды для разработки баз данных используется платформа IBM Lotus Domino/Notes, так как уже существует корпоративная информационная система, разработанная на этой платформе, и накоплены значительные объемы информации для автоматизации переводов [2].

В соответствии с выдвинутыми выше требованиями создается корпоративная сеть, в которую объединены несколько АРМов переводчиков. Платформа Lotus Domino/Notes позволяет использовать не только сервер приложений для формирования и ведения баз данных, содержащих информацию для переводческой деятельности, но и почтовый сервер и Web–сервер для получения и обмена дополнительной информацией помимо имеющейся в корпоративной сети. Сервер Domino поддерживается самыми распространенными операционными системами (ОС), например, такими как Windows, Linux, Solaris, iSeries, AIX, z/OS, что позволяет легко переходить из одной ОС на другую или использовать несколько серверов под разными ОС.

Сервис репликаций позволяет синхронизировать состояние копий баз данных на разных серверах и клиентских машинах (в нашем случае это АРМ переводчика). Система репликаций дает возможность организовать коллективную работу над переводимым документом благодаря функции автоматического управления версиями документа, отслеживающей изменения оригинала документа на всех АРМах.

Платформа Lotus Domino/Notes поддерживает формат Unicode, что позволяет работать с многоязычными документами. Начиная с 7-ой версии, в Lotus Notes встроен текстовый редактор и сервис проверки орфографии. Кроме того, результирующий файл может быть представлен в формате ряда текстовых редакторов, например, Word.

На рис. 1. приведена схема корпоративной сети на базе Lotus Domino/Notes, объединяющая пять АРМов переводчика.

4.2 Комплекс баз данных «Перевод»

На сервере Lotus Domino находится комплекс баз данных «Перевод», в составе которого имеется три базы данных: «Тексты для перевода», «Результаты поиска» и «Память переводчика».

База данных «Память переводчика», содержит специальные электронные библиотеки - словари и параллельные тексты.

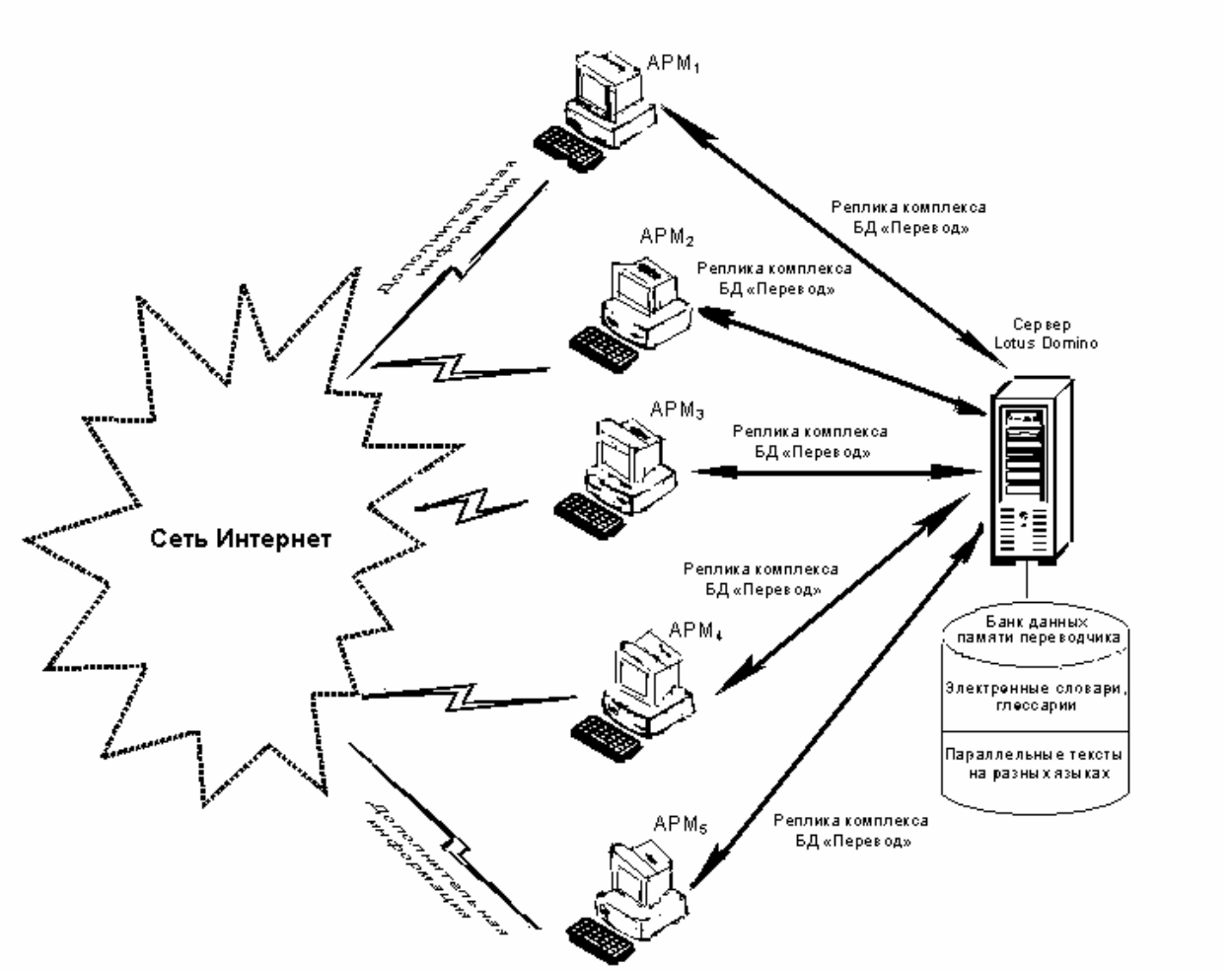

Рис. 1. Схема получения и обмена информацией в корпоративной сети, объединяющей АРМы переводчиков

Библиотека параллельных текстов формируется, в основном, за счет полных текстов документов международного характера (договоров, конвенций, меморандумов, резолюций ООН и т.д.).

Терминологические словари включают лексику из двуязычных глоссариев (русско-английских и русско-французских) по внешнеполитической деятельности, которые составлялись вручную, а также термины из многоязычного электронного словаря по внешней политике [2]. В настоящее время объем словаря составляет около 10 тыс. терминов, однако после перевода в электронную форму материалов на бумажном носителе объем словаря должен значительно вырасти (примерно до 20-25 тыс. терминов).

В базе данных «Тексты для перевода» содержатся исходные документы, предназначенные для перевода, и результирующие (переведенные) документы.

База данных «Результаты поиска» содержит документы, найденные в «Памяти переводчика», в которые входят фрагменты, выделенные переводчиком в исходном тексте.

База данных «Память переводчика», доступна только на сервере, ее нет на АРМах, а реплики остальных баз имеются на каждом из АРМов.

4.3 Схема компонентов АРМ переводчика

На каждом АРМе должны быть установлены следующие средства:

- клиент-серверное программное обеспечение Lotus Notes;
- текстовый редактор Word Microsoft Office 2007;
- система оптического распознавания текстов Abby Fine Reader 8.0;
- электронный переводчик Promt ;
- электронный словарь Lingvo 9.0;
- реплика базы данных «Тексты для перевода»;
- реплика базы данных «Результаты поиска».

4.4 Подготовка текстов для ввода в базу данных «Память переводчика»

При вводе электронных терминологических словарей в базу данных памяти переводчика не возникает трудностей. Процесс осуществляется с помощью стандартных процедур импорта. Ввод параллельных текстов требует дополнительной обработки, которая заключается в выравнивании текстов на уровне абзацев. Такая опция присутствует во многих системах автоматизированного перевода, а также существуют специальные программы выравнивания текстов, доступные в бесплатном пользовании, например, bligner или bitext2tmx.

В нашей системе задача выравнивания значительно упрощается, так как тексты международных документов тщательно выверены и соблюдено полное соответствие их частей (статей, абзацев) на всех рабочих языках.

Как известно, в основе автоматизированной обработки информации на русском языке лежит морфологический анализ. Авторы использовали программу морфологического анализа, разработанную Абрамовым В.Е. на языке Borland C++ Builder 6.

Для подготовки импортируемых файлов разработана специальная программа, выполняющая разбивку параллельных текстов на абзацы и проведение морфологического анализа текста каждого абзаца на русском языке.

При вводе в базу данных для словарей и параллельных текстов предусмотрена одна общая форма, в которой имеются поля для записи названий документов и соответствующих абзацев текстов на двух языках, а также поля для результатов морфологического анализа текстов каждого абзаца.

На рис. 2 показано представление в базе данных «Память переводчика» русско-английского словаря, а на рис. 3 - русско-английских параллельных текстов. База данных содержит библиотеки параллельных текстов и терминологических словарей, сгруппированные для пар языков: русский - английский, русский - французский, русский – испанский.

4.5 Поиск информации в базе данных «Память переводчика»

Lotus Notes предлагает довольно полный объем средств поиска, которые позволяют найти указанные сведения в заголовках документов и в текстах документов по нескольким словам или фразам и с помощью поисковых запросов, составленных на основе булевской логики.

Например, можно составить запрос на поиск слов или словосочетаний, которые обязательно должны присутствовать в документе, кроме того, указать критерий их близости, регистры составляющих их букв, их веса.

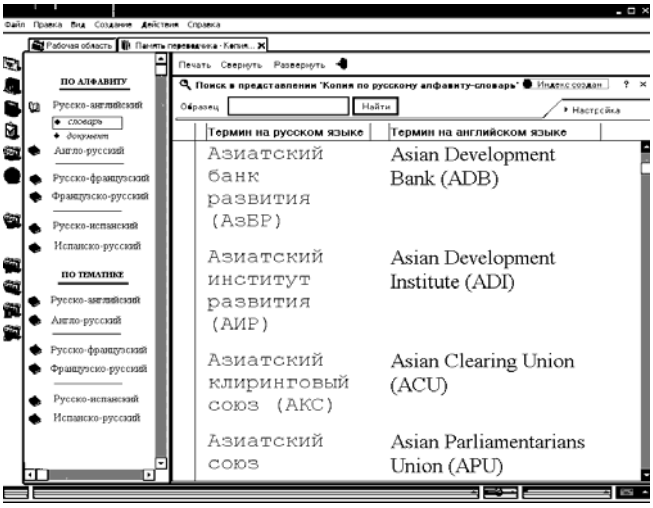

Рис. 2. Представление русско-английского словаря

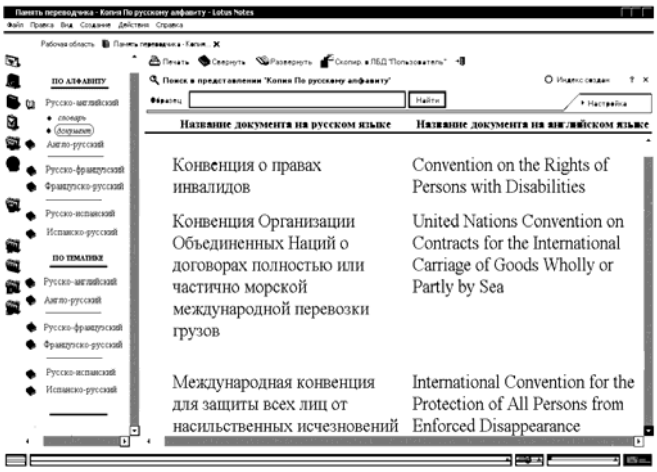

Рис. 3. Представление русско-английских параллельных текстов

Однако для решения поставленной задачи невозможно обойтись только стандартными средствами поиска. Это не слишком удобно для пользователя. Так, открыв найденный в результате поиска документ, нужно путем просмотра документа найти нужный абзац, опираясь на выделенные цветом ключевые слова из запроса. Кроме того, операторы поиска с учетом близости слов в предложении или абзаце работают не всегда корректно в силу того, что в алгоритмах поиска заложен не очень удачный критерий определения границ предложений и абзацев.

Не предусмотрена также операция выделения мышью какого-либо фрагмента текста и поиска его в базе данных.

Стандартные средства не позволяют найти близкие по смыслу к переводимому тексту или абзацу тексты в «памяти переводчика».

Для реализации указанных операций авторы системы разработали специальные программные средства на объектно-ориентированном языке программирования LotusScript .

Обращение из скрипта Lotus к программе морфологического анализа происходит при

выполнении морфологического разбора выделенного фрагмента переводимого текста. Затем в базе данных «Память переводчика» средствами Lotus находятся все документы, отвечающие запросу - текстовой строке, составленной из основ слов фрагмента. В каждом документе, найденном по запросу, ищутся абзацы, содержащие эту сформированную текстовую строку. Сравниваемая строка выбирается из результатов морфологического анализа текста абзаца, полученного на этапе подготовки текста для импорта в базу данных «Память переводчика». Все абзацы, соответствующие запросу, записываются в базу данных «Результаты поиска» вместе со ссылкой на полный текст документа.

4.6 Технология работы с базами данных

Исходные тексты для перевода можно разместить в базе данных различными способами:

- импортировать с помощью средств Lotus Notes;
- ввести с клавиатуры;
- скопировать через буфер обмена из Интернета или какого-либо текстового редактора.

Обязательными для заполнения полями являются поле «наименование», в которое вносится заголовок документа, и поле «исходный текст», куда непосредственно заносится текст документа. В поле «перевод текста» информация заносится в процессе работы. По желанию переводчика там может быть размещен текст, переведенный автоматически с помощью системы машинного перевода. Работая над переводом текста, переводчик может выделить мышью какой-либо фрагмент и выбрать электронную библиотеку (терминологический словарь или корпус параллельных текстов) на нужном языке, в которой будет проводиться поиск, щелкнув по соответствующей пиктограмме на панели действий (см. рис.4).

Результатом выполненного действия будет переход к представлению базы данных «Результаты поиска», в котором указывается текущая дата, фрагмент, по которому проводился поиск и название документа, содержащего этот фрагмент (см. рис. 5). Для удобства пользователей результаты содержат только те абзацы документов, которые включают найденные фрагменты. Документ открывается с помощью щелчка мышью по его названию. Текст располагается в двух колонках, причем каждому абзацу на русском языке соответствует его перевод на иностранном языке, найденный фрагмент выделяется цветом. Под названием документа находится значок гиперссылки, нажав на который переводчик может просмотреть полный текст документа (см. рис. 6).

Если фрагмент найден в терминологическом словаре, то искомый документ представляет собой словарную статью из соответствующего двуязычного словаря.

В том случае, когда фрагмент не найден в базе данных памяти переводчика, в специальном окне об этом выдается сообщение.

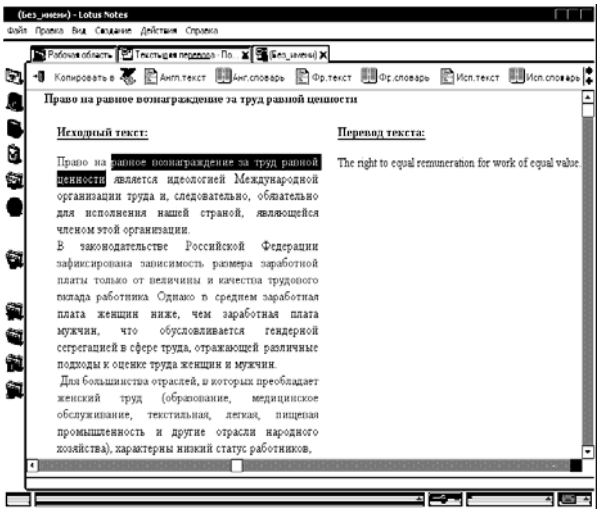

Рис. 4. Представление исходных текстов

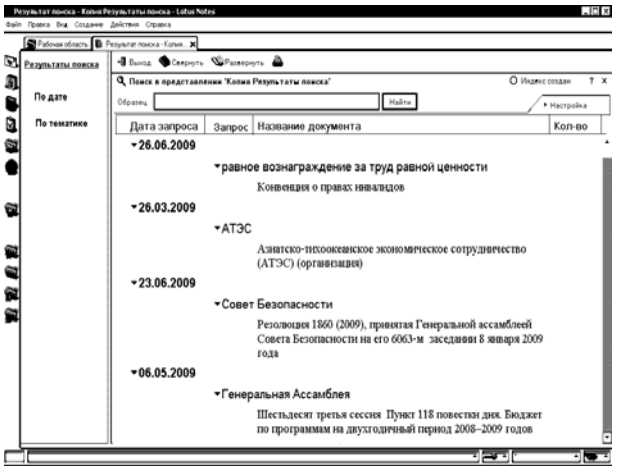

Рис. 5. Результаты поиска переводных эквивалентов

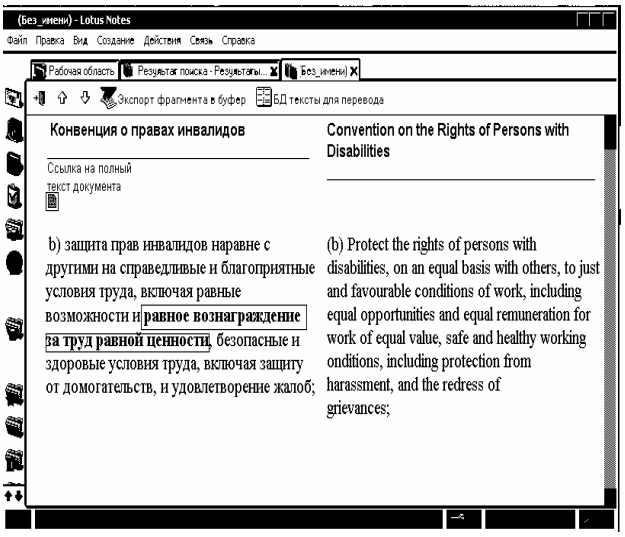

Рис. 6. Абзац из текста документа, содержащий переводные эквиваленты

Переводчик легко может сравнить переводные эквиваленты на двух языках и выделить фрагмент

на иностранном языке, соответствующий искомому фрагменту на русском языке. Нажав на кнопку «БД тексты для перевода», находящуюся на панели действий, он может вернуться в базу данных исходных текстов к документу, который им переводится. Затем в поле «перевод текста» можно вставить выделенный фрагмент на иностранном языке и продолжить процесс перевода.

Пользователь может воспользоваться также стандартными возможностями Lotus Notes, самостоятельно составляя поисковый запрос с учетом близости слов в предложении или абзаце.

5 Заключение

Описанная в данной работе экспериментальная система, разработанная на платформе IBM Lotus Domino/Notes, позволяет эффективно накапливать и осуществлять поиск информации в базе данных памяти переводчика. Пользователь может либо выделить любой фрагмент из переводимого текста, либо составить запрос с использованием языка Lotus Notes и провести поиск в терминологическом словаре или в библиотеке параллельных текстов на любом из рабочих языков.

По мере накопления терминологии и текстов в базе данных «Память переводчика» и следуя пожеланиям переводчиков, можно расширить возможности системы, включив функцию нечеткого поиска, которая имеется в некоторых системах, чтобы искать близкие по смыслу тексты, но допускающие некоторое перефразирование.

С этой целью можно использовать тезаурус. В настоящее время мы располагаем тезаурусом по общественно-политической тематике объемом 10 тысяч понятий. Если в исходном тексте, предназначенном для перевода, и текстах на русском языке из библиотеки параллельных текстов провести поиск слов и словосочетаний из тезауруса и найденные словарные единицы заменить на заглавные дескрипторы тезауруса, то можно (при достаточно хорошем покрытии текстов тезаурусом) решать проблему вариативности терминологии. Затем для каждого абзаца исходного текста можно находить в базе памяти переводчика абзацы, отвечающие некоторому критерию близости текстов, и выдавать пользователю соответствующие им абзацы параллельных текстов на нужном языке.

Однако все доработки системы можно будет проводить после некоторого периода эксплуатации системы.

Литература

[1] Абрамов В.Е. Автоматическое рубрицирование и реферирование текстовой информации (в том числе на иностранных языках). Автореферат дис. канд. техн. наук.- М.: Стандартинформ, 2008, -27 с.

[2] Абрамова Н.Н., Косматова Л.В., Майорова Н.С., Матюшина Н.А., Шелимова И.Н. Многоязычный электронный словарь по внешней политике. – Тез. докл. 6-ой Международ. конф. «НТИ-2002» (Москва, 16-18 октября 2002 г.), 2002. – с.

[3] Белоногов Г. Г., Калинин Ю.П., Хорошилов А.А. Компьютерная лингвистика и перспективные информационные технологии. – М.: Русский мир, 2004. – 248 с.

[4] Русина Л. Технические средства в работе переводчика: сравнительная характеристика. – Материалы регионал. конф. ProZ.com (Харьков, 18-19 октября 2008 г.). http://www.proz.com/conference/68?page=schedule&mode=details&session_id=2144

[5] Силонов А. Программы, помогающие переводчику.- PC Week/RF, N13, 2000. – с. 45. http://www.bntp.ru/home.asp?artId=23.

[6] Nagao M. A framework of a mechanical translation between Japanese and English by analogy principle // A. Elithorn, R. Banerji (eds.), Artificial and Human Intelligence, Edinburgh: North-Holland, 1984. Pp. 173-180.

[7] Интернет-сайт компании SDL Trados: http://www. trados.com

[8] Интернет-сайт компании DejaVu: http://www.atril.com

[9] Интернет-сайт компании SDL International: http://www.sdlintl.com

[10] Программа автоматизации перевода Metatexis: www.metatexis.com

[11] Система автоматизированного перевода Transit: http://www.star-portals.net/Transit/default.aspx

[12] Система автоматизированного перевода OmegaT: http://www.omegat.org/ru/reviews.html

[13] Система автоматизированного перевода Termstar: http://www.star-group.net/star-www/description/termstar/star-group/eng/star.html

[14] TM-программа (накопитель переводов) Wordfisher: www. wordfisher.com/wf4.htm

Corporate translation Network using special digital Libraries

N.N.Abramova, V.E. Abramov, A.A. Karnatskaja, V.M. Rozhkov

The article describes an IBM Lotus Domino/Notes corporate translation network interconnecting several automated translator workstations (AWS). The development of the system was aimed at automating the translation process to meet the modern quality and speed standards of translation. The system is based on the translator memory principles when a great body of manually translated texts in four languages (Russian, English, French and Spanish) is used. Components of the AWS are described and examples of system's functioning are given.