Т. А. Горностай

АРХИТЕКТУРА ЛАТЫШСКО-РУССКОГО МОДУЛЯ СИСТЕМЫ МАШИННОГО ПЕРЕВОДА ДЛЯ ПОДДЕРЖАНИЯ СОЦИАЛЬНОЙ КОММУНИКАЦИИ

Работа представлена кафедрой прикладной лингвистики РГПУ им. А. И. Герцена.
Научный руководитель – доктор филологических наук, профессор Л. Н. Беляева

В статье рассматриваются особенности архитектуры латышско-русского модуля специализированной системы машинного перевода для поддержания социальной коммуникации. Внимание уделено отдельным компонентам архитектуры модуля.

Ключевые слова: *социальная коммуникация, лингвистический автомат, машинный перевод, архитектура системы машинного перевода.*

T. Gornostay

LATVIAN-RUSSIAN MODULE ARCHITECTURE OF THE MACHINE TRANSLATION SYSTEM FOR SOCIAL COMMUNICATION

The article studies the peculiarities of the Latvian-Russian translation module architecture of the specialised machine translation system for social communication. Separate components of the module architecture are described.

Key words: *social communication, linguistic automaton, machine translation, machine translation system architecture.*

Машинный перевод (МП) представляет собой область компьютерной лингвистики, которая занимается теорией и практикой применения компьютера к задаче перевода текста с одного естественного языка на другой. Сегодня МП представляет собой перспективное направление с почти 200 действующими практическими системами МП [9], которые приме-

няются как средство извлечения знаний в научной и технической сфере. Несмотря на общепризнанный факт о невозможности полностью автоматизировать процесс перевода [1; 3; 5], развитие переводческих технологий имеет научную и социально-политическую значимость.

Исторические события 1 мая 2004 г., когда Европейский Союз принял в свои члены десять новых стран-участниц, остро поставили вопрос о необходимости средств поддержки социальной коммуникации в условиях существования обмена информацией на глобальных и национальных языках. Гипотеза исследования состоит в предположении, что специализированная система МП в структуре многофункционального лингвистического автомата может использоваться как средство поддержки социальной коммуникации, позволяя пользователю преодолевать языковой барьер.

Под термином *социальная коммуникация* понимается непрофессиональное общение в отличие от *неофициального общения*. Основной функцией социальной коммуникации является социальная адаптация той части общества, которая не владеет государственным национальным языком. В соответствии с теорией коммуникации систему коммуникации образуют участники коммуникации: источник и получатель информации, а также сама информация и ее преобразователь.

В настоящем исследовании под участниками социальной коммуникации понимаются государственные институции и средства массовой печатной информации, с одной стороны, и человек, не владеющий государственным языком, в данном случае – пользователь системы МП, с другой. При этом информацией являются следующие типы документов: законодательные акты, публицистические издания, а также информация, опубликованная государственными структурами на порталах сети Интернет, а преобразователем информации – система МП.

В статье рассматриваются особенности организации специализированной системы латышско-русского МП для поддержания социальной коммуникации. Именно данное направление перевода имеет особую важность, так как Латвия является страной с почти половиной русскоязычного населения, значительную часть которого составляют жители среднего и старшего возраста, а также мигранты, не владеющие латышским языком на должном уровне, достаточном для обеспечения социального взаимодействия.

Уже существуют практические системы МП, предоставляющие возможности перевода текстов с/на латышский язык [9]. Руководствуясь лингвистическими и кибернетическими принципами построения лингвистических автоматов, принципами модульности архитектуры и открытости системы [2; 4; 6; 7; 8], систему латышско-русского МП для поддержания социальной коммуникации можно рассматривать в качестве модуля действующей системы МП, в нашем случае - системы англо-латышского МП Tildes Tulkotвjs в структуре лингвистического автомата Tildes Birojs, разработанного компанией Tilde.

Таким образом, латышско-русский модуль МП должен быть встроен в базовую систему МП с добавлением необходимых компонентов. Для этого необходим анализ организации базовой системы МП. В результате такого анализа были установлены принципы создания и включения латышско-русского модуля МП, выделены готовые компоненты, которые можно использовать при создании проектируемой системы, и определен набор дополнительных компонентов, необходимых для латышско-русского МП. Рассмотрим архитектуру латышско-русского модуля МП.

В структуру латышско-русского модуля входят следующие компоненты:

- модуль опознания латышского языка;
- модуль фрагментации латышского текста;
- модуль парсинга (машинного анализа), объединяющий морфологический и синтаксический анализ латышского языка;
- модуль синтаксического трансфера (преобразования);
- модуль лексического трансфера (собственно перевод);
- модуль согласования, включающий в себя модуль морфологического анализа русского языка;
- модуль синтеза выходной структуры.

Опознанный латышский текст посылается на обработку модулю фрагментации, целью которого является преобразование исходного текста в цепочку отдельных предложений и словоформ. На следующем этапе проводится морфологический анализ словоформ – определение грамматических характеристик словоформы, приведение всех словоформ к их словарной (канонической) форме и синтаксический анализ предложения с помощью парсера. Парсер разработан на основе универсального CYK-алгоритма, который проводит восходящий синтаксический анализ и позволяет частичный разбор предложения входного языка. Оригинальный CYK-алгоритм, поддерживающий контекстно-свободную грамматику Хомского, дополнен признаками, которые используются в правилах на ограничение, присвоение либо передачу грамматической характеристики другой вершине дерева [10].

На стадии синтаксического трансфера входная конструкция латышского языка трансформируется в выходную конструкцию русского языка с помощью правил. В результате детального анализа синтаксиса текстов на материале источников социальной коммуникации и сопоставительного анализа грамматических и функциональных свойств латышского и русского языков были разработаны следующие виды трансформационных правил: установление связи между двумя или тремя вершинами дерева выходного предложения, изменение порядка слов, скрытие вершины дерева, добавление новой вершины, перенос грамматической информации с одной вершины дерева на другую.

Далее на этапе лексического трансфера осуществляется перевод выходной структуры – извлечение переводных эквивалентов из двуязычного латышско-русского словаря. Автоматический словарь в структуре системы МП является ее центральным компонентом. В структуру системы латышско-русского МП входят три автоматических словаря: словарь основ и парадигм для морфологического анализа латышского языка, двуязычный переводной словарь и словарь основ и парадигм для морфологического анализа русского языка. Лексические единицы в переводном словаре хранятся в их канонической форме. Социальная коммуникация определяет подход к отбору лексики в словарь. Базовым двуязычным словарем для системы МП для поддержания социальной коммуникации является словарь общей лексики, который расширяется за счет лексики, выбранной из источников социальной коммуникации.

На выходе лексического трансфера каждой лексической единице латышского языка присваивается ее русский переводной эквивалент из двуязычного словаря. На данном этапе трудности могут вызвать лексические единицы, проанализированные на этапе морфологического анализа, но не вошедшие в двуязычный словарь, как минимум, по двум причинам. Во-первых, наличие трех автоматических словарей в системе вызывает необходимость координации лексического состава всех словарей, так как эффективность перевода определяется точностью идентификации всех лексических единиц текста. Однако это требует дополнительных ресурсов и неэффективно, с точки зрения экономии времени и усилий разработчика, поэтому для оптимизации процесса перевода наиболее представительным должен быть словарь морфологического анализатора латышского языка для идентификации лексических единиц входного текста. Остальные словари необходимо расширять по возможности.

Во-вторых, это может быть связано с малочастотными лексическими единицами или отдельными словоформами, которые не всегда целесообразно включать в словарь. Для решения этой задачи необходим специальный анализ с помощью дополнительного словарного модуля, в котором хранятся грамматические и словообразовательные правила преобразования. Например, малочастотные лексические единицы, образованные префиксальным способом словообразования с использованием греческих и латинских слов, как правило, не входят в состав двуязычного автоматического словаря. Так, в латышском словаре хранится прилагательное *funkcionāls*, однако нет его деривата *polifunkcionāls*, образованного с исполь-

зованием префикса *poli-*. С помощью разработанных правил для перевода таких слов отдельно переводится префикс *poli-*, а перевод частотного прилагательного *funkcionāls* извлекается из двуязычного словаря. Таким образом, в результате анализа устанавливается перевод лексической единицы *polifunkcionāls*.

Другой пример – анализ неличных форм глагола латышского языка. Для перевода латышского действительного причастия настоящего времени *smaidošs* используется правило, с помощью которого причастие приводится к его канонической глагольной форме – инфинитиву *smaidīt*. Русский переводной эквивалент *улыбаться* извлекается из двуязычного словаря, и ему присваиваются грамматические характеристики латышского причастия, полученные на этапе морфологического анализа. Затем в результате грамматических преобразований с помощью правил согласования синтезируется русское действительное причастие настоящего времени *улыбающийся* на последнем этапе МП.

Следующим является этап согласования, при котором соответствующие морфологические характеристики присваиваются русским лексическим единицам. Далее с помощью разработанных правил согласования решаются две основные задачи. На парадигматическом уровне морфологические характеристики латышской и русской лексических единиц должны быть согласованы и приведены к одному формату, в соответствии с которым будет синтезирована русская выходная лексическая единица. Типичным примером применения правила согласования в данном случае служит правило изменения значения рода существительного. Например, слово *māja* является существительным женского рода в латышском языке, а его переводной эквивалент *дом* – существительным мужского рода. Перед применением правила согласования каждой из лексических единиц приписана информация о значении рода. Однако для успешного синтеза русской лексической единицы позиция морфологической характеристики рода должна быть заполнена значением рода переводного эквивалента. Правило согласования (рис. 1) корректирует данную позицию морфологической характеристики с помощью функции присвоения значения рода *Analyze.Gender* русского переводного эквивалента.

На синтагматическом уровне правила согласования используются для согласования отдельных единиц в цепочке русских словоформ в соответствии с правилами грамматики русского языка. Рисунок 2 демонстрирует правило согласования, в соответствии с которым прилагательное в атрибутивной функции согласуется с существительным в роде, числе и падеже.

Последний этап работы латышско-русского модуля МП – оформление результатов перевода, который представляет собой процедуру порождения словоформ, последовательность и морфологические характеристики которых определены на предыдущих уровнях.

Такой подход к организации латышско-русского модуля МП позволяет создать продукт, дающий возможность человеку, не знающему или плохо знающему латышский язык, извлечь информацию из текста на иностранном языке и поддерживать социальную коммуникацию с ее помощью.

```
Rule(N)// gender
{
        Target.Gender = Analyze.Gender;
}
```

Рис. 1. Пример правила согласования для изменения значения рода

```
Rule(N<-attr-A)// gender, number, case
{
        Child.Gender = Parent.Gender;
        Child.Number = Parent.Number;
        Child.Case = Parent.Case;
}
```

Рис. 2. Пример правила согласования прилагательного с существительным

СПИСОК ЛИТЕРАТУРЫ

1. *Беляева Л. Н.* Лингвистические автоматы // Вестн. С.-Петерб. отд-ния Рос. акад. естеств. наук. 1999. Т. 3. № 1. С. 73–82.

2. *Беляева Л. Н.* Лингвистические автоматы в современных гуманитарных технологиях. СПб.: ООО «Книжный Дом», 2007. 192 с.

3. *Зубова И. И.* Информационные технологии в лингвистике. Минск: Изд-во Мин. гос. лингвист. ун-та, 2001. 211 с.

4. Лингвистические ресурсы автоматизированного рабочего места филолога. СПб.: Изд-во ИнфоДа, 2004. 184 с.

5. *Пиотровский Р. Г.* Лингвистический автомат и его речемыслительное обоснование. Минск: Изд-во Мин. гос. лингвист. ун-та, 1999. 196 с.

6. *Пиотровский Р. Г.* Новые горизонты машинного перевода // Научно-техническая информация. Сер. 2. Информационные процессы и системы. М.: Рос. Акад. Наук, Всерос. Инст. науч. и техн. информ (ВИНИТИ). 2002. № 1. С. 17–29.

7. Работа лингвистического автомата с языками различной типологии // Структурная и прикладная лингвистика. СПб.: Изд-во СПбГУ. 2004. Вып. 6. С. 260–277.

8. *Beliaeva L.* Machine Translation Versus Dictionary and Text // Journ. of Quantitative Linguistics. 2003. Vol. 10. No 2. P. 193–211.

9. Compendium of Translation Software [Electronic resource]: directory of commercial machine transl. systems and computer-aided transl. support tools. 15th ed. Geneve: European Association for Machine Translation. 2008. URL: http://www.hutchinsweb.me.uk/Compendium.htm (05.06.2009). (Файл в формате *pdf, общий объем 129 с.).

10. *Skadiņa I., Vasiļjevs A., Deksne D., Skadiņš R., Goldberga L.* Comprehension Assistant for Languages of Baltic States // Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007. Tartu: University of Tartu, 2007. P. 167–174.

REFERENCES

1. *Belyaeva L. N.* Lingvisticheskiye avtomaty // Vestn. S.-Peterb. otd-niya Ros. akad. yestestv. nauk. 1999. T. 3. N 1. S. 73–82.

2. *Belyaeva L. N.* Lingvisticheskiye avtomaty v sovremennykh gumanitarnykh tekhnologiyakh. SPb.: OOO «Knizhny Dom», 2007. 192 s.

3. *Zubova I. I.* Informatsionnye tekhnologii v lingvistike. Minsk: Izd-vo Min. gos. lingvist. un-ta, 2001. 211 s.

4. Lingvisticheskiye resursy avtomatizirovannogo rabochego mesta filologa. SPb.: Izd-vo InfoDa, 2004. 184 s.

5. *Piotrovsky R. G.* Lingvisticheskiy avtomat i ego rechemyslitel'noye obosnovaniye. Minsk: Izd-vo Min. gos. lingvist. un-ta, 1999. 196 s.

6. *Piotrovsky R. G.* Novye gorizonty mashinnogo perevoda // Nauchno-tekhnicheskaya informatsiya. Ser. 2. Informatsionnye protsessy i sistemy. M.: Ros. Akad. Nauk, Vseros. Inst. nauch. i tekhn. inform (VINITI). 2002. N 1. S. 17–29.

7. Rabota lingvisticheskogo avtomata s yazykami razlichnoy tipologii // Strukturnaya i prikladnaya lingvistika. SPb.: Izd-vo SPbGU. 2004. Vyp. 6. S. 260–277.

8. *Beliaeva L.* Machine Translation Versus Dictionary and Text // Journ. of Quantitative Linguistics. 2003. Vol. 10. No 2. P. 193–211.

9. Compendium of Translation Software [Electronic resource]: directory of commercial machine transl. systems and computer-aided transl. support tools. 15th ed. Geneve: European Association for Machine Translation. 2008. URL: http://www.hutchinsweb.me.uk/Compendium.htm (05.06.2009). (Файл в формате *pdf, общий объем 129 с.).

10. *Skadiņa I., Vasiļjevs A., Deksne D., Skadiņš R., Goldberga L.* Comprehension Assistant for Languages of Baltic States // Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007. Tartu: University of Tartu, 2007. P. 167–174.