УДК 004.9

Ю.В. КРАК, А.В. БАРМАК, С.А. РОМАНИШИН

МЕТОД ОБОБЩЕННЫХ ГРАММАТИЧЕСКИХ КОНСТРУКЦИЙ ДЛЯ АВТОМАТИЗИРОВАННОГО ПЕРЕВОДА ТЕКСТОВОЙ ИНФОРМАЦИИ В ЖЕСТОВЫЕ АНАЛОГИ

Аннотация. Описан подход к решению проблемы перевода с вербального на жестовый язык глухих людей. Для обеспечения перевода построены обобщенные грамматические конструкции на основе корпуса предложений, используемых людьми с недостатками слуха при обычном общении. Рассмотрены практическая реализация инфологических моделей для словарей флективного и жестового языков для работы с данными конструкциями и алгоритм перевода. Получено экспериментальное подтверждение эффективности предложенного подхода на базе украинского языка.

Ключевые слова: автоматизированный перевод, текст, жестовый язык, инфологическая модель.

ВВЕДЕНИЕ

В настоящей работе рассматривается жестовый язык (ЖЯ) — естественный язык, передающий информацию с помощью движения рук и пальцев, выражения лица, положения туловища человека. Этот язык используется как составная часть для общения и служит основным коммуникационным средством для

глухих людей [1, 2]. Жестовые языки не являются визуальной интерпретацией обычных языков, они обладают своей грамматикой, могут использоваться для обсуждения различных тем: от простых и конкретных — до возвышенных или абстрактных. Лексика разговорного ЖЯ еще недостаточно изучена, особенно идиоматика, фразеология, морфология. Автоматическое приписывание разговорного ЖЯ всем формам словесной и письменной речи некорректно. Разговорный ЖЯ глухих людей невозможно рассматривать в традиционных моделях лингвистики. Для изучения морфологии разговорной жестовой речи более подходит описание по принципу: от значения — к форме. В силу этого проблемы, возникающие при построении систем автоматизированного перевода произвольного текста в ЖЯ, заключаются в нахождении взаимооднозначных пар конструкций, которые передают смысл информации. Автоматизированный перевод [3] отличается от машинного [4, 5] тем, что предполагает следующие формы взаимодействия с пользователем: частично автоматизированный перевод (использование человеком-переводчиком компьютерных словарей); системы с разделением труда, в которых компьютер обучен переводить только жестко структурированные фразы (без последующих поправок), а перевод фразы с другой структурой выполняет человек.

Рассмотрим синтаксические особенности ЖЯ на примере трех типовых структур предложений: субъект–объект–глагол; субъект–глагол–объект; глагол–субъект–объект. Подлежащее и сказуемое в таких предложениях связаны предикативно. Предложения с одной предикативной связью назовем простыми предложениями. Отметим, что порядок слов в предложениях с одной предикативной связью в большинстве разговорных языков мира описывается одной из трех указанных типовых структур [6]. В ЖЯ простые предложения служат основным способом коммуникации и делятся на повествовательные, вопросительные и побудительные [7].

Для реализации автоматизированного перевода исходной текстовой информации в ЖЯ предположим, что возможны следующие пары: предложение на исходном языке — аналог на ЖЯ в виде некоторых обобщенных конструкций, построенных, в частности, на простых предложениях. Если проанализировать наборы пар, то можно, зафиксировав порядок следования слов в предложении, получить некоторую обобщенную форму, когда вместо конкретных слов в предложении выписаны наборы слов, которые могут использоваться (находиться) на зафиксированных местах. В результате будет сформирован достаточно небольшой (относительно общего количества предложений некоторого языка) список обобщенных грамматических конструкций для перевода. Такие грамматические конструкции в дальнейшем могут использоваться как шаблоны и правила в системах машинного перевода на ЖЯ [8–10].

Не ограничивая общности, покажем возможности такого подхода на реализации системы перевода на ЖЯ для флективных языков, в частности украинского.

ПОСТАНОВКА ЗАДАЧИ

Сформулируем задачу: разработать и реализовать систему автоматизированного перевода предложений на украинском языке на соответствующие конструкции-предложения украинского ЖЯ.

Предлагается решение поставленной задачи со следующими ограничениями:

1) учитываются только простые предложения;

2) смысл ЖЯ ограничивается фиксированным перечнем тем и ситуаций;

3) система переводит (без последующих поправок) только обученные, жестко заданные структуры фраз и предложений и не искажает смысла;

4) постоянное расширение (обучение) перечня структур перевода.

ИНФОЛОГИЧЕСКАЯ МОДЕЛЬ СЛОВАРЕЙ УКРАИНСКОГО И ЖЕСТОВОГО ЯЗЫКОВ

Для достижения цели перевода необходим грамматический словарь украинского языка. Украинский язык флективный, т.е. синтетического типа, в котором доминирует словоизменение с помощью флексий — формантов, сочетающих не-

сколько значений. Флективное строение языка противоположно агглютинативному, в котором каждый формант имеет только одно значение. Для разработки грамматического словаря используем теорию лексикографических систем [11–13].

Построение грамматического словаря флективного языка определяется наличием формальной модели словоизменения. Это означает установление и формализацию лингвистических критериев, согласно которым все множество слов языка разбивается на определенные подмножества, взаимное пересечение которых пусто, а внутри каждого из них словоизменение происходит по одинаковым правилам. Подмножества слов с такими свойствами называются словоизменительными парадигматическими типами.

Под парадигматическими типами будем понимать группу лексем, словоизменительная парадигма которых характеризуется одинаковым количеством грамматических форм. Внутри группы словоизменение происходит по одному и тому же (единственному) правилу. Украинский язык относится к аналитико-синтетическому типу: слова, принадлежащие к одному парадигматическому классу, имеют одинаковые флексии в соответствующих грамматических значениях и одинаковый характер чередования в основе; соответствующие аналитические формы строятся по одинаковым моделям их образования.

Украинскому языку присущи следующие грамматические категории, определяющие словоизменение: s — число (единственное (s_1), множественное (s_2)); g — род (мужской (g_1), женский (g_2), средний (g_3), превосходная степень (g_4), мужской или средний род для одушевленного предмета (g_5), женский или средний род для одушевленного предмета (g_6), мужской или женский род для одушевленного предмета (g_7), неизменяемая словарная единица (g_8), сравнительная степень (g_9); i — обобщенная категория, включающая падеж, лицо, время, состояние, способ (именительный (i_1), родительный (i_2), дательный (i_3), винительный (i_4), творительный (i_5), предложный (i_6), звательный (i_7), инфинитив (i_8), первое лицо, повелительное наклонение (i_9), второе лицо, повелительное наклонение (i_{10}), третье лицо, повелительное наклонение (i_{11}), первое лицо, будущее время (i_{12}), второе лицо, будущее время (i_{13}), третье лицо, будущее время (i_{14}), первое лицо, настоящее (будущее) время (i_{15}), второе лицо, настоящее (будущее) время (i_{16}), третье лицо, настоящее (будущее) время (i_{17}), первое лицо, настоящее время (i_{18}), второе лицо, настоящее время (i_{19}), третье лицо, настоящее время (i_{20}), действительное причастие, настоящее время (i_{21}), деепричастие, настоящее время (i_{22}), прошедшее время (i_{23}), действительное причастие, прошедшее время (i_{24}), страдательное причастие, прошедшее время (i_{25}), безличная форма, прошедшее время (i_{26}), деепричастие, прошедшее время (i_{27}), неизменяемая единица (i_{28})).

Каждое слово отнесем к следующим классам (частям речи) (p): существительные (p_1), числительные порядковые (p_2), числительные количественные (p_3), числительные (p_4), числительные типа «два» (p_5), глаголы несовершенного и совершенного вида (p_6), глаголы совершенного вида (p_7), глаголы несовершенного вида (p_8), причастия (p_9), прилагательные (p_{10}), местоимения (p_{11}), местоимения-прилагательные (p_{12}), наречия (p_{13}), междометия (p_{14}), союзы (p_{15}), частицы (p_{16}), предлоги (p_{17}), слова сказуемого (p_{18}), вводные слова (p_{19}), аббревиатуры (p_{20}), связки (p_{21}), деепричастия (p_{22}), личные местоимения (p_{23}), союзы и частицы (p_{24}), наречия и частицы (p_{25}), числительные с предлогом (p_{26}), существительные с предлогом (p_{27}), деепричастия совершенного вида (p_{28}), местоимения с предлогом (p_{29}), деепричастия несовершенного вида (p_{30}).

Для украинского языка введем парадигматические типы (T), грамматические классы, грамматические категории (табл. 1).

Исходя из того, что украинский язык флективный (грамматические значения передаются флексиями), слова языка моделируются в виде комбинации неизменной и переменной составляющих:

$$x = c(x) \& f(x), \tag{1}$$

где $c(x)$ — часть лексемы x, которая в процессе словоизменения остается неизменной (квазиоснова), $f(x)$ — ее переменная составляющая (квазифлексия), & — конкатенация.

Таблица 1

Парадигматический тип	Грамматический класс	Грамматическая категория, определяющая словоизменение	Количество грамматических значений
T^1	p_1	$s * (i_1, \dots, i_7)$	14
T^2	$p_2 \cup p_4 \cup p_9 \cup p_{10} \cup p_{12}$	$(g_1, g_2, g_3) * ((s_1, s_2) * (i_1, \dots, i_7))$	24
T^3	$p_3 \cup p_{11}$	$(i_1, \dots, i_6)$	6
T^4	p_5	$(g_1, g_2) * (i_1, \dots, i_6)$	12
T^5	$p_6 \cup p_8$	i_8, $(i_{10} * s_1), (i_9 * s_2), (i_{10} * s_2), (i_{12} * s_1)$, $((i_{12}, \dots, i_{17}) * s), i_{21}, i_{22}$, $(i_{23} * (g_1, g_2, g_3) * s_1)$, $(i_{23} * s_2), i_{24}, i_{25}, i_{26}, i_{27}$	26
T^6	p_7	$i_8, (i_{10} * s_1), (i_9, s_2), (i_{10}, s_2)$, $((i_{12}, i_{13}, i_{14}) * s)$, $(i_{23} * (g_1, g_2, g_3) * s_1)$, $(i_{23} * s_2), i_{24}, i_{25}, i_{26}, i_{27}$	18
T^0	$\bigcup_{j=13}^{30} p_j$	i_{28}	1

Реализация такой модели хранения слова излишне избыточна, поскольку в украинском языке около двух миллионов слов (во всех словоизменениях) и только около ста тысяч слов-инфинитивов, т.е. вместо полного текста слова в таблице хранится номер слова-инфинитива из соответствующего множества инфинитивов, номер позиции в слове, в которой слово-инфинитив неизменно, и номер флексии (переменной части слова) из соответствующего множества.

Таким образом, слова языка представляются в виде

$$W = \{W_i : W_i = \{I_{i_1} \in I, F_{i_2} \in F, k, g \in G, s \in S, in \in In\}\}, \tag{2}$$

где W_i — параметры слова ($i = 0, \dots, N-1$, N — количество слов в словаре), F — множество всех возможных окончаний слов (флексий), k — номер позиции в слове-инфинитиве, с которой начинает конкатенироваться флексия (возможны случаи, когда $k = 0$ для словоформы, полностью отличной от инфинитива), $G = \{g_1, \dots, g_9\}$, $S = \{s_1, s_2\}$, $In = \{i_1, \dots, i_{28}\}$, I — множество слов-инфинитивов украинского языка (для глаголов — инфинитивы, для существительных — слова в именительном падеже, единственном числе и т.д.):

$$I = \{I_i : I_i = \{wordinf,\ p \in P\}\}, \tag{3}$$

где $P \in \{p_1, \dots, p_{30}\}$, $wordinf$ — слово-инфинитив.

Формирование множеств W, F, I происходит следующим образом (для всех слов-инфинитивов языка):

- k = длина (общая часть для слова-инфинитива и всех его словоформ);
- для всех словоформ определяем окончание (отличающаяся часть) и добавляем его к множеству F (если такого окончания в нем нет);
- добавляем слово-инфинитив к множеству I;
- формируем множество W согласно (2).

Прямой задачей для модели (2) является воспроизведение полного текста слова по трем параметрам — I_{i_1}, F_{i_2}, k, реализуемое соответствующим оператором H:

$$H = \begin{cases} Left(I_{i_1}, k-1) \& F_{i_2}, & k \neq 0, \\ Left(I_{i_1}, 0) \& F_{i_2}, & k = 0, \end{cases} \tag{4}$$

где *Left*(*Word*, *Length*) — функция получения первых символов (*Length*) из слова (*Word*).

Некоторая сложность возникает для обратной задачи (H^{-1}): для существующего слова найти три параметра, которые его однозначно определяют. Реализация оператора (4) для базы данных в виде недетерминистской функции приводит к тому, что ее невозможно эффективно проиндексировать. Поиск по неиндексированному полю достаточно затратный по времени. Для решения этой проблемы предложен следующий алгоритм. Входное слово разбивается на возможные комбинации неизменной и переменной составляющих, и поиск осуществляется по проиндексированным полям таблиц-множеств инфинитивов (I) и флексий (F):

$$H^{-1} = \begin{cases} i = 0,\ \ l = \text{длина}\,(Word), \\ \text{повторять, пока}\ \ i < l, \\ \quad \textbf{Head} = LEFT\,(Word, i), \\ \quad \textbf{Tail} = RIGHT\,(Word, l - i); \\ \quad \text{если существует инфинитив}\ I_{i_1} \in I, \\ \quad \text{который начинается на}\ \textbf{Head}, \\ \quad \text{и существует флексия}\ F_{i_2} \in F, \text{равная}\ \textbf{Tail}, \\ \quad \text{то определяем}\ (I_{i_1}, F_{i_2}, k); \\ \quad \text{или}\ \ i = i + 1 \\ \text{Конец цикла} \end{cases} \tag{5}$$

Для ЖЯ структура словаря в связи с отсутствием в ней словоизменений несколько проще. Инфологическая модель украинского языка и ЖЯ приведена на рис. 1.

В блоке «Словарь украинского языка» содержится множество слов украинского языка (W), в блоках «Флексии» и «Инфинитивы» — соответственно множества флексий (F) и инфинитивов (I). Для каждого слова хранятся ссылки на грамматические категории, определяющие словоизменение и часть речи. Множество жестов находится в блоке «Словарь жестов» и содержит ссылки на слово, которым обозначается данный жест, и тематику, в которой он чаще всего используется.

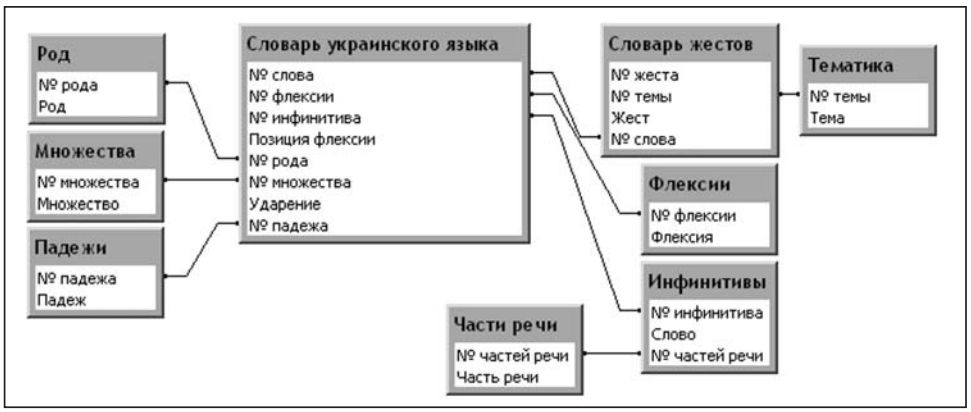

Рис. 1. Инфологическая модель словарей украинского и жестового языков

ИНФОЛОГИЧЕСКАЯ МОДЕЛЬ ОТНОШЕНИЙ ОБОБЩАЮЩИХ КОНСТРУКЦИЙ ДЛЯ ПЕРЕВОДА

Рассмотрим только простые предложения. Считаем, что любое сложное предложение можно представить как композицию простых предложений. Как для украинского языка, так и для ЖЯ ограничимся следующими типами простых предложений: повествовательные (утвердительные, восклицательные и отрицательные), вопросительные и побудительные.

После составления словарей украинского и жестового языков необходимо построить соответствие между конструкциями предложений украинского языка и ЖЯ. Для этого сформировано множество предложений, взятых из программы-комплекса «Украинский жестовый язык» [14], которая используется в учебных заведениях для глухих людей для овладения жестовой речью. Предложения в модели данных объединены в структуры, полученные путем обобщения. Структуры предложений вместо слов содержат их последовательности, каждая из которых может использоваться при построении предложения. Последовательности могут включать как отдельные слова, так и различные множества. Например, предложение «он идет», «она идет», «время идет» объединены в одну последовательность «{он, она, время, кто, ...} идет». Кроме того, последовательности могут содержать категории словоизменения (p, s, g, i), пересечение которых определяет множество слов.

Структуры предложений представлены в виде таблиц и связей между ними (рис. 2).

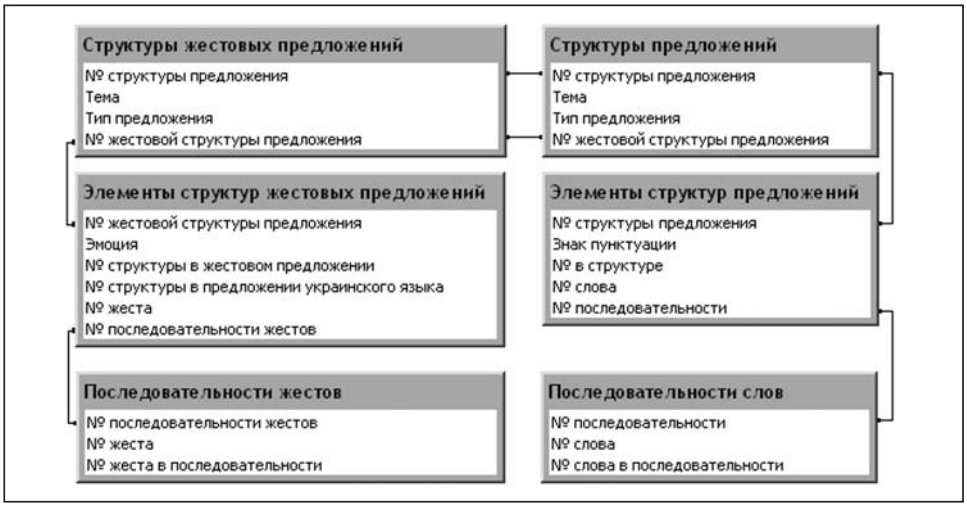

Рис. 2. Структуры текстовых и жестовых предложений

Структуры предложений хранятся в блоке «Структуры предложений». Элементами структур могут быть как отдельные слова, так и последовательности слов. При переводе важно знать порядок элементов структуры предложения и жестового предложения. Для сохранения порядка элементов структуры предложения каждый элемент имеет порядковый номер. «Структуры жестовых предложений» также могут содержать жесты и последовательности жестов. Для обеспечения перевода необходимо для входного предложения найти подходящую конструкцию ЖЯ и жесты, соответствующие каждому слову предложения.

После заполнения базы данных структурами предложений украинского языка были созданы соответствующие им структуры ЖЯ. Для каждого элемента структуры предложений украинского языка зафиксирован соответствующий элемент структуры жестовых предложений. Элементы структур ЖЯ также содержат порядковые номера элементов структур украинского языка для обеспечения соответствия между элементами.

Из множества структур предложений получено множество обобщенных конструкций украинского языка и ЖЯ, которые содержат только категории словоизменения, а не отдельные слова или множества слов. Для каждого элемента структуры предложения выделено множество категории словоизменения (p, s, g, i) и порядок данного элемента в конструкции. Создание обобщенных конструкций позволило на основе анализа категорий словоизменения (p, s, g, i) слов в предложении и порядкового номера элемента получать соответствующую ему структуру жестового предложения. Кроме определения структуры предложения для перевода необходимо соответствие слов жестам.

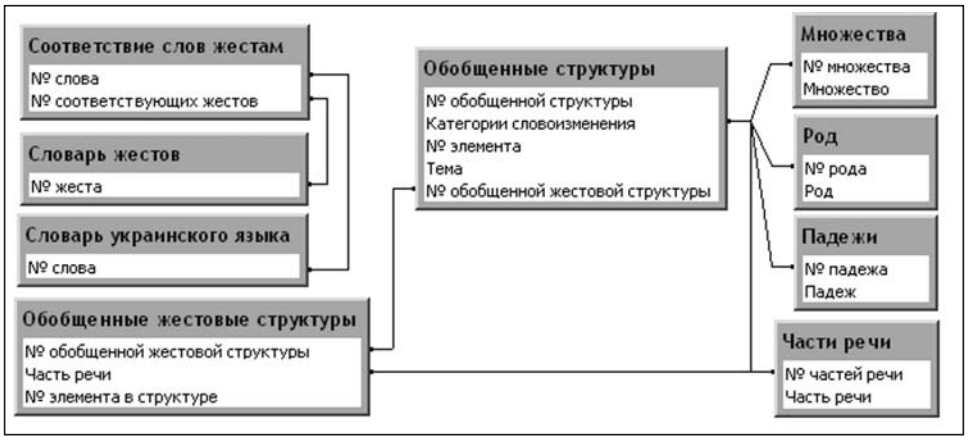

Рис. 3. Инфологическая модель соотношений обобщающих конструкций для перевода

На основе соответствия между структурами предложений и структурами жестовых предложений выделено множество соответствий «слово → жест» (рис. 3).

Для автоматизированного перевода текста с украинского языка на жестовый используется следующая алгоритмическая схема.

1. Предложение поступает на вход, определяются категории словоизменения для каждого слова предложения и осуществляется поиск обобщенной конструкции украинского языка, соответствующей данной структуре.

2. Осуществляется поиск соответствий «слово → жест» для каждого слова входного предложения.

3. Для обобщенной конструкции украинского языка находится соответствующая конструкция ЖЯ, на основе соответствия «слово → жест» выводится результат.

4. При наличии нескольких конструкций ЖЯ предлагаются все возможные варианты перевода. Если не найдено жестов, отвечающих определенным словам предложения, перевод выполняется частично, исключая данные слова. В таком случае предоставляется возможность добавить соответствие «слово → жест».

5. Если обобщенной конструкции для введенного предложения не существует, предпринимается попытка прогноза перевода на основе (p, s, g, i) и порядка слов части входного предложения. В случае корректного перевода эта конструкция добавляется к уже существующим. Если перевод не удовлетворителен, предоставляется возможность добавить новую обобщенную конструкцию языка и соответствующую ей конструкцию ЖЯ.

ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ И ПРОВЕРКА ЭФФЕКТИВНОСТИ ТЕХНОЛОГИИ

В структуру базы данных вносится множество предложений, полученных из программы [14]. Предполагается, что оно покрывает большое количество ситуаций, возникающих в повседневной жизни глухих людей. Данное множество обобщено в структуры предложений, для которых перевод идентичен. С помощью специалистов, преподающих ЖЯ, выполнен перевод структур предложений на ЖЯ и созданы аналогичные конструкции для ЖЯ (рис. 4). С использованием структур предложений украинского языка и соответствующих им предложений ЖЯ сформированы обобщенные конструкции предложений на основе категорий словоизменения (p, s, g, i) уже существующих конструкций. Обобщенные конструкции позволили осуществлять перевод не только для предложений, полностью соответствующих заданным конструкциям, а и на основе анализа структуры категорий словоизменения входящего предложения. Кроме того, из структур предложений украинского языка и ЖЯ выделены соотношения между словами и жестами с учетом тематики предложения. В случае наличия разных вариантов перевода слова переводчику предоставляется возможность выбора из нескольких вариантов наиболее корректного.

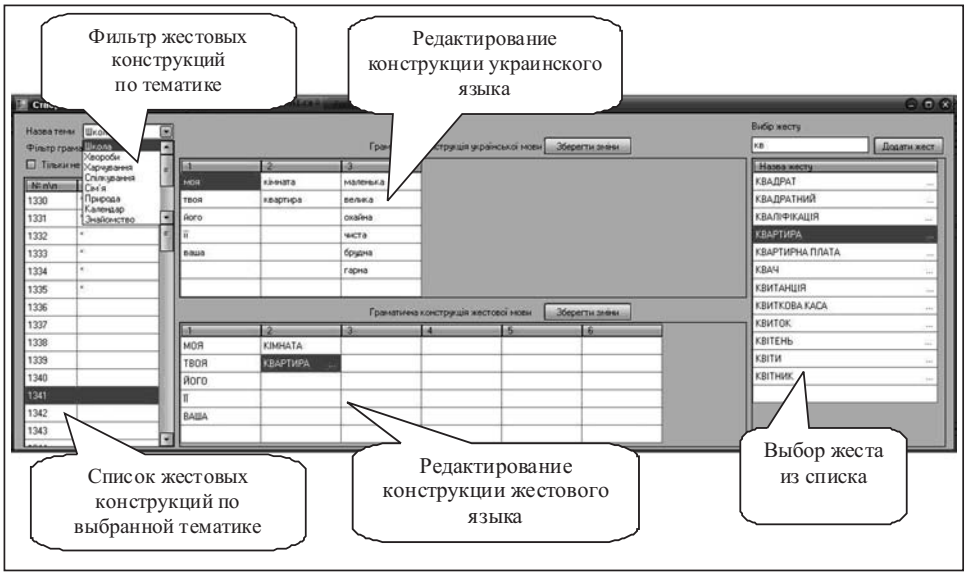

Рис. 4. Модуль создания грамматических конструкций украинского и жестового языков

Для проверки эффективности разработанной технологии использовано множество из 9800 предложений, взятых из программы изучения украинского ЖЯ для специализированных школ. Предложения содержат жесты из тем, часто используемых глухими людьми для повседневного общения. В результате перевода данных предложений и их анализа получено 293 обобщенных конструкции перевода. При тестировании ЖЯ на выбранных предложениях получен однозначный перевод без искажения смысла в 100% случаев. Успешным также был перевод с использованием других предложений с тем же словарным запасом.

ЗАКЛЮЧЕНИЕ

Для обеспечения автоматизированного перевода с украинского языка на ЖЯ построены модели словарей украинского и жестового языков. На их основе создано множество конструкций для двух языков, позволяющих выполнять перевод фиксированного множества предложений, используемых глухими людьми в повседневной жизни. Полученные конструкции обобщены для перевода с учетом структуры категорий словоизменения элементов предложений.

Дальнейшие исследования направлены на разработку на основе предложенной информационной технологии приложений для перевода с украинского языка на жестовый с использованием web-технологий, обеспечение возможности редактирования существующих конструкций и создания новых (уполномоченными лицами), тестирование технологии большим количеством предложений для выявления новых грамматических конструкций и добавления их в базу данных, сбор статистики применения тех или иных конструкций для перевода.

Исследуются также проблемы использования предложенного подхода для других языков.

СПИСОК ЛИТЕРАТУРЫ

1. Stokoe W.C. Sign language structure: An outline of the visual communication systems of the american deaf // Studies in Linguistics. — 1960. — N 8. — 78 p.
2. Зайцева Г.Л. Жестовая речь. Дактилология: Учеб. для студ. высш. учеб. заведений. — М.: Гуманит. изд. центр ВЛАДОС, 2000. — 192 с.
3. Bowker L. Computer-aided translation technology: a practical introduction. — Ottawa: Univ. of Ottawa Press, 2002. — 185 p. — http://www.google.com/books?id=ly29-mc6dO0C&printsec=frontcover&hl=uk#v=onepage&q&f=false.

4. Madsen M.W. The limits of machine translation. — Copenhagen: Center for Language Technology, Univ. of Copenhagen, 2009. — 116 p.
5. Tomlin R.S. Basic word order. Fundamental principles. — London: Croom Helm, 1986. — 308 p.
6. Адамюк Н.Б., Чепчина І.І. Синтаксичні особливості української жестової мови: на прикладі простого речення // Жестова мова і сучасність. — 2009. — № 4. — С. 170–191.
7. Lavie A., Denkowski M.J. The METEOR metric for automatic evaluation of machine translation // Machine Transl. — 2009. — **23**, Issue 2–3. — P. 105–115.
8. Morrissey S. Assessing three representation methods for sign language machine translation and evaluation // Proc. 15th Intern. Conf. Eur. Assoc. for Machine Transl. — Leuven (Belgium), 2011. — P. 137–144.
9. Stein D., Schmidt C., Ney H. Sign language machine translation overkill // Intern. Workshop on Spoken Language Translation. — Paris, 2010. — P. 337–344.
10. Baldassarri S., Royo-Santas F. Trends on human-computer interaction: research, development, new tools and methods. — London: Springer-Verlag, 2009. — 165 p.
11. Широков В.А. Феноменологія лексикографічних систем — К.: Наук. думка, 2004. — 327 с.
12. Корпусна лінгвістика: Монографія / В.А. Широков, О.В. Бугаков, Т.О. Грязнухіна та ін. // Укр. мовно-інформ. фонд НАН України. — К.: Довіра, 2005. — 472 с.
13. Любченко Т.П. Програмно-технологічні аспекти створення граматичних лексикографічних систем // Проблеми програмування. — 2007. — № 3. — С. 61–75.
14. Программа-комплекс «Українська жестова мова». — http://www.mon.gov.ua/education/average/programs_gluh.

Поступила 06.08.2013