УДК 651.926:681.3

Н.М. Мищенко, О.Д. Фелижанко, Н.Н. Щёголева

Универсальная система программ обработки тематических текстов

Предложена технология построения языковых процессоров для обработки тематических текстов на естественных языках, состоящих каждый из двух частей: универсальной программной, ориентированной на класс языков, и информационной в виде машинного представления грамматики и схемы перевода конкретных языков. Рассмотрен язык описания информационной части, по которому генерируется машинное представление.

A technology is presented for constructing universal language processors for the texts in natural languages. Each processor is composed of two parts: a universal part oriented to a class of languages and an information part containing the machine representation of grammar and translation methods for a specific language. A specification language for the information part allowing for generating its machine representation is suggested.

Запропоновано технологію побудови мовних процесорів для обробляння тематичних текстів натуральними мовами, кожний з яких складається з двох частин: універсальної програмної, орієнтованої на клас мов, та інформаційної, що містить машинне представлення граматики і схем перекладу конкретних мов. Розглянуто мову опису інформаційної складової, за яким генерується її машинне представлення.

Введение. Глобальная компьютерная сеть сделала возможным распространение идей и средств, способствующих научно-техническому прогрессу. Большое значение приобретают средства общения на естественных языках разноязычных субъектов, в том числе и с компьютером, ставшим активным участником диалога.

Проблема общения с компьютерами на естественном языке имеет свою историю. Первой попыткой наладить такое общение была реализация на ЭВМ «Киев» идеи В.М. Глушкова обучать машину распознавать смысл представленных ей фраз на естественном языке. Эта идея была темой успешного выступления В.М. Глушкова в 1962 году на конгрессе *IFIP*–62 (*International Federation for Information Processing*) в Мюнхене [1]. Через несколько лет появились диалоговые системы общения на формальных языках, предпосылки появления которых изложены в статье [2]. С лингвистической точки зрения проблема общения с компьютером рассматривается в монографии [3], где приводятся примеры различных моделей типа «текст–смысл» и проблемы их практической релизации. В последнее десятилетие получили распространение онтологии (классификация знаний). На их основе создана *Semantic Web* (Семантическая паутина). Однако прогресс в этой сфере, похоже, мало повлиял на процесс общения человека с богатством Интернета. Конкретная же задача – как передавать смысл сообщения на естественных языках, если в каждом языке существует синонимия не только лексическая, но и выраженная неявно другими языковыми средствами, а также как машине сформулировать ответ – решается в каждом отдельном случае индивидуально. Для общения с компьютером можно использовать подмножества естественных языков или строить язык–посредник (формальный или полуформальный) между человеком и компьютером. Во всех возможных вариантах общения непременно возникает задача построения программ перевода с естественного на язык общения.

В статье предлагается технология построения универсальных программ для пословной обработки тематических текстов на примере языковых процессоров *DUET* – для перевода и *FEST* – для лексико-статистического анализа текстов. В данной реализации слово – это также составные слова и устойчивые словосочетания. Пословная обработка научно-технических текстов возможна благодаря практически отсутствующей омонимии слов, достаточно простому синтаксису и ограниченному словарю употребляемой лексики в таких текстах.

Характеристика подхода

Особенность предложенных языковых процессоров (ЯП) *DUET* и *FEST* состоит в том, что язык входных текстов обоих ЯП является их параметром, а сами они ориентированы на специалистов–нелингвистов.

Приводим основные свойства выбранной сферы применения ЯП *DUET* и *FEST*.

• Пословная обработка текстов – сравнительно проста в реализации и такова, что допускает формальное описание перевода. Здесь уместно напомнить о так называемом правиле 20/80 [4]: если с помощью средств автоматизации, на которые затрачено 20% ресурсов, решается 80% задачи, то расходы на средства экономически целесообразны. Перевод научных текстов с помощью ЯП *DUET* вполне приемлем, поскольку более 85% фраз в переводах не нуждаются в исправлении.

• Поскольку в профессиональной деятельности пользователя–нелингвиста лингвистическая система имеет второстепенное значение, для него более важно – удобство в пользовании, а не полнота системы, которая непременно ведет к ее усложнению. Кроме того, многие лингвистические процессоры, напоминают «черные ящики» и не содержат средств их приспособления к потребностям пользователей из конкретной области знаний [5]. В связи с этим и возникла потребность построения мобильной программы перевода *DUET*, поддерживаемой средствами автоматизации построения словарей, которые формировал бы пользователь для своей специальности.

Ориентация ЯП на профессиональный язык позволяет повысить эффективность работы ЯП за счет словарей сравнительно небольшого объема, какими пользуется та или иная область тематических знаний. Ибо каким бы полным не был словарь (и, пропорционально, высокой цена соответствующей лингвистический системы), всегда может случиться, что российский биологический термин «древесные побеги» переводится на украинский язык, как «дерев'яні втечі».

• Построение словарей – процесс сложный и требует определенных лингвистических знаний. При настройке ЯП на язык конкретной специальности предлагаем придерживаться так называемой текстовой идеологии, суть которой заключается в том, что слова для занесения в словарь выбираются из профессиональных текстов пользователя на основе частотных списков лексики, которые генерирует ЯП *FEST*. Такие словари лучше всего отражают активную лексику текстов определенной профессии и не содержат лишних слов. Лексико-статистические исследования текстов показали, что профессиональная лексика имеет самую высокую частоту употребления среди полнозначной лексики.

• Основная программа в ЯП *DUET* и *FEST* – морфологический анализ (МА), распознающий слова. Пословный перевод в значительной степени компенсируется подобием синтаксических структур переводимых языков из класса допустимых за исключением нескольких конструкций, трансформирующихся в процессе перевода. Редактирование перевода таких конструкций ложится на пользователя.

Описание технологии построения ЯП *DUET* и *FEST*

Намереваясь построить ЯП, программы которого способны обслуживать различные профессиональные языки, необходимо предусмотреть удобный переход от одной пары языков к другой. Такое требование к построению ЯП привело к понятию: входной и выходной языки – это параметры ЯП *DUET*, и таковым же есть входной язык для ЯП *FEST*. Параметризация ЯП *DUET* и *FEST* означает деление каждого из них на две части: универсальную (программы) и переменную информационную часть (параметр).

Универсальная часть – это общие для разноязычных входных текстов программы, выполняющие: морфологический анализ, проверку согласования соседних словоформ во входном тексте, формирование перевода, списков распознанных или нераспознанных словоформ. Созданные для генерации первого ЯП, они в дальнейшем дополняются соответствующими информационными составляющими, образуя конкретные ЯП по заказу пользователей. Многократное использование универсальных программных компонентов повышает их надежность. Они недоступны для изменения пользователями.

В отличие от невидимых для пользователя программ ЯП видимыми для них являются: метаязыки *Lmorf* и *Lduet* для создания спецификаций, соответственно, *Smorf* и *Sduet* информационных составляющих ЯП, а также программы, объединенные названием «генератор *DUAL*», генерирующие параметр – информационную составляющую на основе спецификаций *Smorf*

и *Sduet*. Итак, переход от одного языка к другому (при переводе – от одной пары языков к другой) выполняется заменой в структурах данных ЯП информационной составляющей на сгенерированную генератором *DUAL* по спецификациям другого языка.

Архитектуру генератора *DUAL* для генерации МП *DUET* составляют компоненты:

{*Lmorf*, *GENtbl*, *Lduet*, *CONdic*, *GENw*},

где *Lmorf* – метаязык описания морфологической информации входного и выходного языков;

GENtbl – программа–генератор морфологических таблиц *Mtbl* входного языка и *Mtbl*–1 выходного по их описаниям, соответственно, *Smorf* и *Smorf1* на метаязыке *Lmorf*;

Lduet – формальный декларативный метаязык для спецификации *Sduet* лексики входного и выходного языков и схем перевода;

GENw – генератор словоформ входного и выходного языков по спецификации *Sduet*;

CONdic – генератор информации *D* для ЯП *DUET* по описанию *Sduet*.

Спецификация лексики – ответственная работа, результат которой целесообразно проверить перед генерированием на ее основе информационной составляющей будущего ЯП. Такая проверка проводится по результатам работы программы *GENw*.

После проверки и исправления ошибок в спецификации *Sduet* программа *CONdic* генерирует информационную часть *D*, составляющую словари входного и выходного языков, содержащие служебные и неизменяемые слова, а также основы изменяемых слов. Все они снабжены морфологической информацией. Каждая статья входного словаря содержит ссылку на соответствующую статью в выходном словаре. Таким образом фиксируется схема перевода.

Подаем структуру МП *DUET* в терминах составляющих.

{*T*, *Mtbl*, *Mtbl*–1, *D*},

где *T* – универсальная программная составляющая ЯП *DUET*.

Генераторы *GENw* и *CONdic* реализованы с помощью РСП «Терем» [6], поскольку входные тексты для них принадлежат к классу контекстно-свободных языков.

Спецификации информационной составляющей ЯП *DUET*

Спецификации для ЯП *DUET* – это описание морфологической информации: *Smorf* входного и *Smorf1* выходного языков на метаязыке *Lmorf*, а также описание *Sduet* лексики входного языка и схемы ее перевода на метаязыке *Lduet*. Следует отметить, что язык *Lmorf* служит описанием морфологической информации и для ЯП *FEST*.

Спецификация морфологической информации. Спецификация морфологической информации входного языка *Smorf* для обоих ЯП *DUET* и *FEST* и выходного *Smorf1* для ЯП *DUET* содержит элементы трех типов: объекты собственно *языка* (алфавит, окончания), объекты *метаязыка* (названия падежей, лиц, классов лексем и т.д.) и *системные коды* объектов метаязыка.

Предполагается, что словоформа изменяемой лексики входного или выходного (при переводе) языка состоит максимум из трех частей: основы – обязательной начальной части словоформы, которая не изменяется при склонении или спряжении, возможно, суффикса и окончания. Окончания совпадают с каноническими. Суффикс – это часть словоформы, которая находится между основой и окончанием. Если суффикс общий для всех словоформ одной и той же лексемы, его следует отнести к основе. Суффикс может состоять из нескольких канонических суффиксов или не совпадать ни с одним из них. Списки суффиксов и основ, в отличие от окончаний, не подаются отдельно, а формируются генератором *CONdic* в процессе обработки спецификаций лексики *Sduet*. Префикс всегда относим к основанию.

Переходим к определению понятий метаязыка *Lmorf*. Последовательность окончаний языка, которые принимаются некоторым классом словоформ, назовем *кортежем окончаний*, определяющим этот класс. Для именных частей языка каждое окончание кортежа соответствует определенному падежу, начиная с именительного единственного числа и оканчивая предложным множественного числа. Для глаголов – это окончания в лицах единственного и множественного числа. Каждый класс словоформ

получает уникальное мнемоническое имя – *шифр* класса, который является названием и соответствующего кортежа окончаний. Таким образом, спецификация морфологической информации входного языка *Smorf* и выходного *Smorf*1 состоит из трех разделов:

1) алфавит языка анализируемых текстов;

2) список падежей, лиц и шифров, где каждый элемент сопровождается присвоенным ему числовым системным кодом и коротким описанием его грамматических значений;

3) шифры с кортежами соответствующих окончаний.

На основе спецификаций *Smorf* и *Smorf1* программа *GENtbl* строит морфологические таблицы, соответственно, *Mtbl* и *Mtbl*–1, каждая из которых содержит списки указанных ранее объектов, древовидное представление окончаний, список омонимов окончаний, представление кортежей окончаний в виде двумерного массива чисел – адресов окончаний в списке окончаний. Подробнее генерация морфологических таблиц изложена в [7].

Отметим, что спецификация морфологической информации украинского языка составляет 29 Кб, а соответствующие морфологические таблицы 33 Кб, аналогичные показатели и для русского языка, в то время как для английского (аналитического) – 3 Кб и 2 Кб соответственно,.

Программа морфологического анализа (МА) включена в оба ЯП – *DUET* и *FEST*. Анализ слова программой МА выполняется с двух концов по очереди. После некоторого числа шагов получаем результат в виде слова или основы слова, возможно, суффикса и окончания или сообщения, что слово в словаре не найдено. Что делать дальше с этими данными, зависит от параметров программы МА и информации, сопровождающей найденное слово в словаре, в частности, ссылки на слово–перевод и на его грамматические признаки. Изложение алгоритма МА приведено в [8].

Разработка и реализация алгоритма МА для ЯП *DUET* и *FEST* проводились с учетом опыта реализации алгоритма МА Мельчука И.А. [9]. Именно в этой работе изложена идея деления алгоритмов МА на универсальные программы, т.е. общие для определенного класса языков, и таблицы, содержащие морфологическую информацию конкретного языка из класса допустимых. Эта идея применяется авторами статьи для ЯП *DUET*, где, кроме морфологической информации, в качестве параметра подается и описание схемы перевода.

Спецификация пословного перевода. Составление спецификации перевода – творческая работа, поскольку требует исследования способов перевода каждого слова, а не бездумного занесения в компьютер слов из бумажного словаря. Проще всего составлять словарь служебных и неизменяемых слов. Это можно сделать в первую очередь, поскольку такие слова употребляются в текстах одинаково часто и независимо от специальности. Их количество не превышает тысячи–двух, а в текстах на флективных языках их употребление достигает 30% по отношению ко всей лексике текста. В текстах на английском языке их часть составляет до 48%.

Спецификация пословного перевода естественного языка на метаязыке *Ldic* – это последовательность правил, разделенных символом ';'. Каждое правило содержит неизменяемое слово или общую основу нескольких словоформ, или целое словосочетание и, если необходимо, грамматическую информацию для анализа входного и синтеза выходного текста. Рассмотрим несколько правил перевода лексики с русского языка на украинский.

1. Если переводится неизменяемое или составное слово, то описание перевода исчерпывается указанием выходного слова:

часто => *; *всегда* => *завжди*; *чаще всего* => *найчастіше*.

Звездочка в первом примере на месте перевода означает совпадение перевода со словом.

2. Для перевода чаще всего требуется дополнительная информация в виде схемы перевода, отделенной от выходной цепочки двоеточием. Эту информацию составляют шифры кортежей окончаний, отметки падежей или лиц и суффиксы. Обычно схема перевода состоит из двух частей, разделенных знаком «=». Информация слева от знака равенства касается входной цепочки, справа – выходной.

множеств => *множин*: *ісо* = *іжа*_1.

Схема перевода означает: все словоформы существительного русского языка среднего рода с основой *множеств*- принимают окончания из кортежа, имеющего шифр *ісо* и переводятся на украинский язык существительными женского рода с основой *множин*- и с окончаниями из кортежа *іжа*_1.

3. Суффикс, который употребляется не во всех словоформах с одной и той же основой присоединяется к основе, как это сделано в правиле:

готовност => *готовн*: *іжь* = *іж*зь_1 «*іст*» (*он*, *озн*, *оо*) «*ост*».

Чередование гласных в суффиксах украинских словоформ обусловило перечень падежей, в которых употребляется суффикс -*іст*- (именительный, винительный, творительный единственного числа) -*ост*- в остальных падежах.

4. Следующее правило предлагает перевод словоформ, образованных от двух основных: *имеющийся* (перевод: *наявний*) и *имеющий* (перевод: *що має*). Указанные входные слова имеют общую основу. Во входном словаре основы не должны совпадать, поэтому в данном случае предлагаются в одном правиле два альтернативных перевода, разделенных знаком '!', каждый с отдельной схемой. Схема первого перевода содержит три шифра, поскольку род прилагательных есть изменяемым признаком, а причастие склоняется как прилагательное:

имеющ => наявн : пмдщий_1/пждщая_1/псдщее_1 = пчий_1/пжа_1/псе_1 ! «що ма» : пмдщий/пждщая/псдщее = «є» (одн) «ють».

Второй перевод (после знака '!') сопровождается схемой без шифра в правой части правила, потому что имеющейся информации достаточно для синтеза выходной цепочки, так как все формы причастия *имеющий* в единственном числе переводятся «*що має*», а во множественном – «*що мають*». Окончания здесь используются как псевдосуффиксы, цепочка (*одн*) означает, что '*є*' присоединяется во всех падежах единственного числа. В остальных падежах присоединяется '*ють*'.

Анализ результатов перевода с русского языка на украинский научных текстов из области проектирования ЭВМ показал, что более 85% слов переведены правильно (не считая слов, отсутствующих в словаре), что свидетельствует о целесообразности применения формального описания пословного перевода. Остальная часть входного текста (менее 15%) не всегда может быть переведена правильно. Одна из причин связана с локальностью области согласования словоформ. Некоторые из согласованных словосочетаний входного текста после перевода становятся несогласованными, в частности, когда слова согласованных словосочетаний находятся в разных предложениях или, будучи в одном предложении, разделены словоформами, не принадлежащими словосочетанию. Не учтены также в ЯП все случаи изменения структуры выходного текста в сравнении со структурой входного. В этом случае придерживаемся принципа: лучше недоделать, чем сделать неправильно. Поэтому редактирование результата перевода человеком необходимо.

Пример русско-украинского перевода. Приведем предложение текста *Rus.txt* (27 Кб) на русском языке, в котором формулы заменены символом '#'. Его перевод на украинский язык с помощью МП *DUET* показывает типичные недостатки перевода.

В тактируемых триггерах, *кроме информационных и управляющих входов*, *есть входы, по которым поступают тактирующие сигналы #, а также установочные входы # и # для принудительной установки триггера в нулевое и, соответственно*, *единичное состояние.*

Перевод. *У*, *що тактуються тригерах, крім інформаційних і керуючих входів, є входи,* (*за,по*) *як*(*-ими, -их*) *поступають тактуючі сигнали* #, *а також установочні входи # і # для примусового встановлення тригера в нульове і*, *відповідно*, *одиничний стан.*

В первой строке перевода слово *тригерах* следует поменять местами с цепочкой *що тактуються,* в другой – оставить предлог *по* и выбрать окончание -*их,* в третьей строке следует согласовать имя прилагательное *нульове* с именем существительным *стан.* Согласование не было выполнено во время перевода, поскольку эти слова во входном тексте разделены другими словами.

Лексико-статистические исследования текстов (МП *FEST*)

Просмотр больших массивов текстов (статей, отчетов) человеком с целью исследования их лексики крайне неэффективен. Простейший способ повышения эффективности исследования – сужение информации для просмотра в виде частотных списков слов и словосочетаний текстов, которые строит ЯП *FEST* ([10]). Для достижения поставленных целей в ЯП *FEST* предусмотрено представление лексики текстов в двух спецификациях: *Sfest*1 – служебной и неизменяемой лексики, *Sfest*2 – полнозначной лексики. Соответственно, генерируются два словаря: *d*1 и *d*2, составляющие словарь *D*1.

Архитектуру генератора *DUAL* для генерации МП *FEST* составляют компоненты:

{*Lmorf*, *GENtbl*; *Lfest*, *CONdic*,
GENw; *FREQlis*, *GENspc* },

где *Lmorf* – метаязык описания морфологической информации входного языка *Smorf*, по которому *GENtbl* формирует морфологические таблицы *Mtbl*;

Lfest – формальный метаязык для описания лексики *Sfest*, состоящей из двух частей, как было сказано выше. В описании лексики присутствуют только левые части правил перевода ЯП *DUET*;

GENw – генератор словоформ по спецификации *Sfest*;

CONdic – генератор информации *D*1, состоящей из двух частей *d*1 и *d*2, по описанию лексики *Sfest*;

FREQlis – программа построения частотных списков лексики текстов;

GENspc – генератор спецификации лексики по результатам МА входного текста.

Таким образом, МП *FEST* – это пятерка:

{*T*1, *Mtbl*, *D*1, *FREQlis*, *GENspc*},

где *T*1 – программная составляющая – морфологический анализ текста.

Рассмотрим две проблемы, которые можно решать с помощью ЯП *FEST* путем лексико-статистической обработки профессиональных текстов: генерация спецификаций лексики, отсутствующей в словаре ЯП и определение тематики текста.

Для этого ЯП *FEST* выполняет функции:

- морфологический анализ (МА) словоформ текста, в процессе которого накапливаются списки нераспознанных или распознанных словоформ по мере их встречи в тексте. Результат зависит от заданного пользователем параметра *S*: если $S = 1$, результат анализа – список найденных слов, если $S = 2$, результат – список ненайденных слов;
- формирование частотных списков распознанных или нераспознанных слов;
- для каждого распознанного существительного поиск (в пределах предложения) согласованных с ним словоформ для составления частотного списка словосочетаний;
- генерация спецификаций неизвестных слов по частотным спискам таких слов.

Приведем примеры правил спецификации слов русского языка для ЯП *FEST*:

часто => *; *в то же время как* => *;
множеств => *: *ico; готовност* => *: *iжь*.

***Расширение словаря полнозначной лексикой со словарем d*1**. Параметр $S = 2$. Лексико-статистический анализ текстов в ЯП *FEST* с целью расширения словаря полнозначной лексики новой, в том числе и профессиональной, базируется на результатах МА текста со словарем *d*1 служебной и неизменяемой лексики, что позволяет выделить полнозначную лексику в список неизвестных слов для любого текста, в котором используется лексика словаря *d*1.

Неизвестные слова в списке представлены основами и окончаниями. Последние могут быть ошибочными, но при высокой частоте употребления их в разных грамматических формах такие ошибки существенно не влияют на результат. Суффиксы не рассматриваются.

Полученный список основ неизвестных слов подается на вход программе *FREQlis* для формирования частотного списка, содержащего основы и в скобках разные окончания всех словоформ с данной основой. В частотном списке основы упорядочены по убыванию частоты вхождения соответствующих словоформ в текст. Предлагаем фрагмент частотного списка основ словоформ русскоязычного текста по лингвистике.

1) 96 3.37% 5 *спис* (*-ов*, *-и*, *-е*, *-0*, *-а*, *-у*, *-ах*, *-ом*, *-ам*, *-ами*);

2) 68 2.49% 1 *текст* (*-ов*, *-ах*, *-ам*, *-а*, -0, *-е*, *-ом*, *-ы*);

3) 67 2.45% 9 *словар* (*-я*, *-ей*, *-и*, *-е*, *-ем*, *-ь*, *-ями*, *-ю*, *-ях*);

4) 61 2.16% 5 *частотн* (*-ых*, *-ые*, *-ый*, *-ом*, *-ого*, *-ому*, *-ым*, *-ыми*);

5) 49 1.74% 9 *термин* (*-ов*, *-ах*, *-ы*, *-ами*).

В каждой строке частотного списка первое число – порядковый номер строки, второе – количество вхождений словоформ с основой в текст, третье – процент вхождений по отношению к общему числу словоформ, четвертое – строка текста, где впервые встретилась словоформа с данной основой. Такой список является входным для программы *GENspc*, работающей в диалоге с пользователем. Алгоритм ее работы заключается в поиске для каждой основы кортежа окончаний, содержащей все окончания, связанные с основой. Если окончаний достаточно для получения однозначного ответа в виде шифра кортежа, то формируется спецификация основы с найденным шифром. Если окончаний недостаточно, то предлагается несколько кортежей для выбора пользователем единственно правильного. Результат – по частотному списку неизвестных слов, содержащем вышеприведенный фрагмент, построена спецификация слов и расширен словарь *d*2. Затем проанализирован тот же текст со словарем *d*2 и параметром $S = 1$. По списку найденных слов построен частотный список, начальный фрагмент которого имеет вид:

1) 96 3.42% 5 *список*

2) 68 2.44% 1 *текст*

3) 67 2.44% 9 *словарь*

4) 61 2.16% 5 *частотный*

5) 49 1.76% 9 *термин*

Формирование спецификаций по списку неизвестных слов в диалоге с компьютером – средство автоматизации составления словарей для ЯП *FEST* ([11]). Для ЯП *DUAL* полученные спецификации пользователю ЯП следует расширить информацией для перевода в выходной язык.

***Определение тематики текстов со словарем d*1**. Параметр $S = 2$. В этом случае, как и в предыдущем, в процессе МА со словарем служебной лексики *d*1 строится список ненайденной в словаре полнозначной лексики, по которому программа *FREQlis* строит частотный список ненайденной лексики по уменьшению частоты её употребления. Экспериментально проверено: термины возглавляют частотный список [10]. Однако более достоверную информацию дают словосочетания с участием высокочастотной лексики, которые можно получить, выполнив шаги, описанные в предыдущем подразделе. Получив расширенный словарь *d*2 за счет слов высокой частоты употребления, можно выполнить МА со словарями *d*1 и *d*2 и с параметром $S = 1$. В результате будет сформирован список найденных слов, а на его основе построен частотный список словосочетаний, объединенных согласованием в числе и падеже с существительным (считаем существительное главным словом любого словосочетания–термина). Те из них, которые содержат слова из верхней части исходного частотного списка незнакомых слов, и будут терминами.

Пример 1. В тексте *Rus*.txt (27 Кб) общая служебная и неизменяемая лексика составляет 43% от всех словоформ. Частотный список известных слов по тексту *Rus.txt* (Фрагмент 1_*Rus*) позволяет сделать вывод о том, что текст относится к вычислительной технике.

Фрагм. 1_*Rus. Известные слова*

N	*част.*	%	*строка*	*словоформа*
1)	106	3.343	13	*регистры*
2)	75	2.365	27	*триггеров*
3)	57	1.798	83	*операциями*
4)	47	1.482	40	*сигналов*
5)	42	1.325	83	*входов*
6)	38	1.198	77	*разряд*
7)	23	0.725	82	*термов*
8)	23	0.725	270	*регистрограмма*
9)	20	0.631	99	*табл*
10)	18	0.631	4	*проектирования*

Специализация текста уточняется частотным списком словосочетаний Фрагм. 2_*Rus*. со словарем, расширенным терминами из фрагмента Фрагм. 1_*Rus*.

Фрагм. 2_*Rus*.

N	*част.*	*стр.*	*словосочетания*
1)	1	9	*задача логического проектирования компонентов*
2)	1	46	*формальной методики проектирования регистров*
3)	1	396	*проектировании схем триггеров регистра*
4)	1	634	*автоматизации проектирования дискретных устройств*

5)	1	70	*методики формального синтеза регистра*
6)	1	72	*структура проектируемого регистра*
7)	9	140	*разряд регистра*
8)	8	113	*информационным входам*

Частотный список словосочетаний упорядочивается сначала по длине словосочетаний (по количеству слов), а затем по частоте употребления: чем короче словосочетание, тем выше частота употребления.

Пример 2. Рассмотрим начальный фрагмент частотного списка, полученного в результате анализа текста на английском языке (конституция Японии) со словарем служебных слов и с параметром $S = 1$.

В тексте *Eng.txt* (33 Кб) служебная лексика составляет 46% словоупотреблений (всего 80 различных служебных слов). В начальном фрагменте частотного списка (Фрагм.1_*Eng*) полнозначных слов в первой колонке – порядковый номер в списке, во второй – количество употреблений слова, в третьей – процент употребленных слов к общему их количеству в тексте, в четвертой – строка текста, где впервые встретилось слово, в пятой – слово.

Фрагм. 1_*Eng*. Известные слова

N част.	%	*строка*	*слово*
1) 164	3.198	11	*shall*
2) 107	2.086	45	*Article*
3) 76	1.482	52	*House*
4) 66	1.287	20	*laws*
5) 42	0.819	30	*right*
6) 41	0.799	8	*Diet*
8) 35	0.682	7	*peaple*
9) 33	0.643	8	*Representatives*
10) 32	0.624	46	*state*
11).30	0.585	54	*Cabinet*
12) 27	0.526	1	*Constitution*

Текст содержит 5129 словоупотреблений, 1200 различных слов. Наибольшую частоту использования имеет слово *shall* (должен, должны), поскольку в основном законе перечисляются обязанности всех ветвей власти и граждан страны.

Заключение. Использование сгенерированных ЯП свидетельствует о целесообразности применения формальных средств спецификации процессов пословной обработки профессиональных текстов. Используя генератор *DUAL*, разработчик словарей освобожден от рутинного труда по его формированию в виде структур данных ЭВМ, вместо этого основное его внимание сосредоточено на создании текстовой формальной спецификации лексики и грамматики, что есть творческая работа, которую невозможно передать автомату. Текстовые спецификации удобно читать, настраивать, хранить в персональных библиотеках в виде файлов, создавать из них нужные композиции машинных словарей.

1. *Glushkov V.M*. Certain Questions of the Theory of Machine Self-learning // Proc. IFIP Congr. – Munich, 1962. – P. 480–481.
2. *Глушков В.М*. Диалог с вычислительной машиной: современные возможности и перспективы // УСиМ. – 1974. – № 1. – С. 3–7.
3. *Попов Э.В*. Общение с ЭВМ на естественном языке. – М.: Наука, 1982. – 360 с.
4. *Krueger C.W*. Software reuse // ACM Computing Serveys: ACM Press. – 1992. – **24**. – N 2. – P. 131–183.
5. *Щоголева Н.М*., *Міщенко Н.М*., *Феліжанко О.Д*. Особливості перекладу українською наукових текстів з інженерії програмування // Проблеми програмування (Матеріали 6-ї міжн. конф. УкрПРОГ'2008, 27–29 трав. 2008 р., Київ), 2008. – С. 261–269.
6. *Мищенко Н.М*. Средства расширения входных языков РСП ТЕРЕМ и их применение // УСиМ. – 1990. – № 5. – С. 55–62.
7. *Міщенко Н.М*. Про засоби генерації програм морфологічного аналізу // Зб. наук. пр. конф. «Людина. Комп'ютер. Комунікація» (Львів, 5–7 трав. 2010 р.). Видво НУ «Львів. політехніка», 2010. – С. 77–80.
8. *Міщенко Н.М*. Система програм морфологічного аналізу науково-технічних текстів // Зб. «Наукові записки». Матеріали П'ятої міжнар. н-п конф. «Мови і світ: дослідження та викладання». Серія: Філологічні науки. – Кіровоград, Ред.-вид. від. КДПУ. – 2011. – 95(2). – С. 538–542.
9. *Мельчук И.А*. Морфологический анализ при машинном переводе (преимущественно на материале русского языка) // Проблемы кибернетики. – 1961. – **6**. – С. 207–276.
10. *Мищенко Н.М*., *Щёголева Н.Н*. О лексико-статистическом анализе научно-технических текстов // KDS 2003. Proc. X-th Int. Conf. (June 16–26, 2003, Varna (Bulgaria). FOI-Commerce, Sofia, 2003. – P. 315–321.
11. *Міщенко Н.М*., *Феліжанко О.Д*., *Щоголева Н.М*. Засоби розширення граматичного словника за частотним списком невідомих слів // Зб. «Наукові записки». Матеріали П'ятої міжнар. н-п конф. «Мови і світ: дослідження та викладання». Серія: Філологічні науки. – Кіровоград, Ред.-вид. від. КДПУ. – 2011. – 95(2). – С. 543–547.

Тел. для справок: +38 044 526-0253, +38 044 243-0240, +38 044 450-4617 (Киев)
E-mail: *nadmykh@ukr.net*, *fel_olga@voliacable.com*, *nat@incyb.kiev.ua*