УНИВЕРСАЛЬНАЯ СИСТЕМА ПРОГРАММ МОРФОЛОГИЧЕСКОГО АНАЛИЗА НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ НА ФЛЕКТИВНЫХ И АГГЛЮТИНАТИВНЫХ ЯЗЫКАХ

Надежда Мищенко

Аннотация. *В настоящей работе рассматривается система программ FEST, предназначенная для морфологического анализа научно-технических текстов на флективных и агглютинативных языках. Систему FEST составляют универсальная программа собственно морфологического анализа MORPH и несколько вспомогательных программ, генерирующих информационную составляющую программы MORPH − морфологические таблицы и словарь конкретного входного языка из допустимого класса языков. Генерация морфологических таблиц и словарей для программы MORPH осуществляется автоматически по формальному описанию морфологии и лексики, выполненному человеком – знатоком языка. Формальное описание лексики может быть частично автоматизировано. Стратегия анализа смешанная: поочередно слева направо − справа налево. Словарь входного языка состоит из неизменяемых слов и основ словоформ (а не лексем или словоформ) и представляет собой несколько словарей, каждый из которых содержит основы одинаковой длины. Каждая основа в соответствующем словаре сопровождается грамматической информацией, позволяющей распознавать все словоформы входного текста с общей словарной основой. Стратегия анализа, структура морфологических таблиц и словарей позволяют выполнять морфологический анализ всех словоформ с заданными в словаре основами.*

Abstract. *This paper describes the software package FEST, which includes a universal program for morphological analysis of scientific and technical texts, MORPH, and several other programs generating data for MORPH. This data includes the morphological tables of a specific input language belonging to the permissible class of inflectional and agglutinative languages and a dictionary. The programs included in the FEST package generate the input language data for the MORPH program using formal descriptions of morphology and vocabularies created by a human expert who knows the language. The analysis strategy is based on an alternation of left-to-right and right-to-left analysis order. The dictionary of the input language contains stems rather than lexemes or word-forms, and consists of several dictionaries, each containing stems of the same length. The stems in the dictionary are accompanied by the grammar information, allowing all the word-forms of the input text to be recognized. The analysis strategy, the structure of the morphological tables and vocabularies enable morphological analysis of all word-forms with stems from dictionary.*

Ключевые слова: *система программ морфологического анализа, спецификация морфологии, спецификация лексики, генерация морфологических таблиц, генерация словарей, результат морфологического анализа (описание и пример).*

Key words: *the software package for morphological analysis, formal descriptions of morphology, formal descriptions of lexemes, morphological tables generation, vocabularies generation, results of morphological analysis (description and example).*

ACM Classification Keywords*: I.2.7. Natural Language Processing – Text analysis.*

Введение

Морфологический анализ (МА) является непременной составной частью многих процессов обработки текстов на естественных языках: перевода, статистических исследований, формирования корпусов текстов и др. В нашем подходе к реализации программы МА MORPH для анализа текстов на флективных и агглютинативных языках использованы две идеи известного лингвиста Игоря Александровича Мельчука из его статьи [Мельчук, 1961]. Первая идея − деление алгоритма МА на две части: программируемую, общую для класса входных языков (универсальный базис программы MORPH), и информационную, содержащую грамматическую информацию о конкретном входном языке (морфологические таблицы) и словарь. Вторая идея касается стратегии выполнения анализа: в качестве наиболее подходящей стратегии МА предлагается смешанная стратегия: поочередно слева направо − справа налево. При этом в алгоритме МА должны быть предусмотрены все возможные разборы каждой словоформы.

Отметим, что программа MORPH ориентирована на использование грамматического словаря, который содержит неизменяемые слова и основы словоформ с сопутствующей грамматической информацией.

Универсальность базиса программы MORPH означает также общность компьютерных структур морфологических таблиц и словарей для всех входных языков. Эти структуры заполняются автоматически специальными программами системы FEST на основе формальных текстовых спецификаций морфологии и лексики, составляемых человеком – знатоком языка. Такой подход к реализации информационного обеспечения программы MORPH имеет ряд преимуществ: во-первых, отпадает необходимость человеку, составляющему спецификацию лексики, знать компьютерную структуру данных, поскольку генерация таблиц и словарей для конкретного входного языка выполняется автоматически; во-вторых, отладка информации в структурах данных, ее изменения выполняются на уровне спецификаций – текстового представления языковых объектов; в-третьих, мобильность информационной части программы MORPH позволяет пользователю самостоятельно генерировать новую программу МА путем автоматической замены имеющейся грамматической информации по спецификации морфологии и лексики другого языка. В частности, программа MORPH анализирует тексты на русском, украинском и турецком (агглютинативном) языках, используя соответствующие таблицы и словари.

Автор не располагает данными об аналогичных подходах к МА научно-технических текстов.

В статье рассмотрены схема взаимодействия программ системы FEST в процессе настройки программы MORPH на конкретный язык, формальные спецификации грамматики и лексики, описание результатов МА и пример анализа предложения на русском языке с предоставлением грамматических признаков слов.

Схема настройки программы МА на конкретный входной язык

Настройка программы МА MORPH на конкретный входной язык осуществляется программами системы FEST, схема взаимодействия которых в процессе построения морфологических таблиц и словарей

представлена на Рис.1, где в овальных рамках поданы имена файлов с информацией, прямоугольные рамки содержат имена программ. Двойные стрелки указывают на информацию, получаемую от пользователя: спецификации морфологии и лексики, соответственно, файлы morph.grm и lex0.spc, а также текст, подлежащий анализу, в файле text.txt. Одинарные стрелки, входящие в прямоугольники, подают входную информацию, исходящие – выходную.

Начальный шаг наиболее ответственный – составление человеком спецификации морфологии (файл morph.grm). По спецификации morph.grm программа GENmorph выполняет генерацию морфологических таблиц (файл morph. tbl), которые используются остальными программами системы FEST.

Следующий шаг: составление спецификации лексики lex0.spc. На этом шаге используется спецификация морфологии morph.grm: каждой основе или неизменяемому слову сопоставляются присущие им постоянные признаки, закодированные в morph.grm и суффиксы, если они участвуют в словоизменении.

Правильность спецификации lex0.spc проверяется на следующем шаге программой GENw с использованием таблиц morph.tbl путем генерации на основе спецификации lex0.spc всех словоформ с заданными в спецификации основами. Правильность полученных словоформ в текущей версии программы MORPH проверяется визуально. Исправленная спецификация – файл lex.spc.

Следующий шаг: спецификация lex.spc программой GENdic преобразуется в словарь D, который является множеством словарей. Каждый словарь содержит лексемы и основы словоформ одинаковой длины, а также связанную с ними грамматическую информацию. После генерации табличной и словарной информации программа MORPH может выполнять морфологический анализ текста (файл text.txt), результат которого формируется в файле text.wrd.

Следует отметить, что указанные расширения .grm, .spc, .txt в названиях исходных файлов со спецификациями морфологии и лексики, а также текста для анализа обязательны, а основные имена могут быть произвольными. Например, файл со спецификацией морфологии русского языка можно назвать rmorph.grm, тогда получим таблицы rmorph.tbl.

Форма результата МА может зависеть от дальнейшего плана действий над ним, поэтому она регулируется специальными параметрами программы MORPH. Например, можно потребовать в качестве результата выдать список слов входного текста, отсутствующих в словаре, либо список найденных слов в нормальной форме, либо список найденных слов входного текста в сопровождении присущих им постоянных и непостоянных грамматических признаков и леммы.

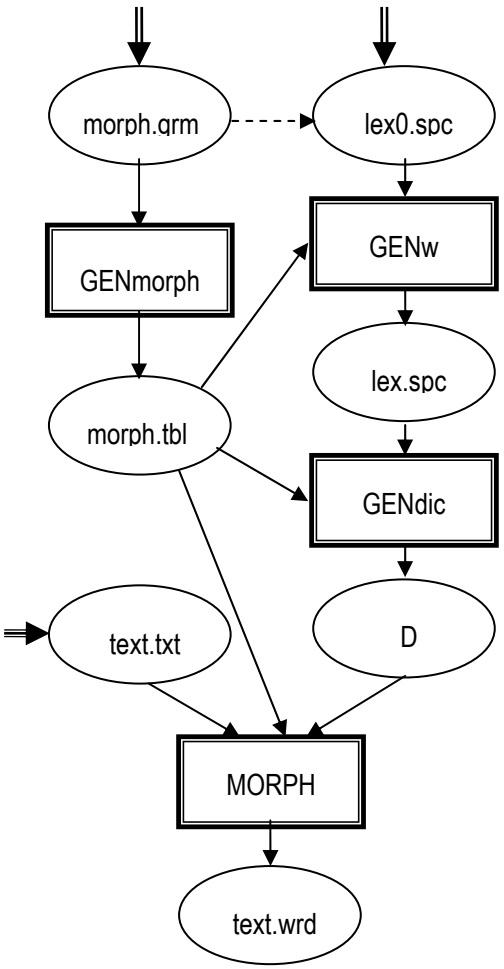

Рис.1. Схема взаимодействия программ системы FEST

Спецификация морфологии

Программа MORPH построена на допущении, что словоформа входного языка состоит не более чем из трех частей: основы − непустой начальной неизменяемой части словоформы, суффикса и окончания. Окончания в основном совпадают с каноническими. Если словоформа в том или ином падеже или лице не имеет окончания, то полагаем, что она имеет нулевое окончание, обозначаемое в спецификации морфологии цифрой 0. Суффикс – это часть словоформы, которая находится между основой и окончанием. Если суффикс общий для всех словоформ некоторой лексемы и к тому же не изменяется во время словоизменения, то его присоединяем к основе. Суффикс может состоять из нескольких канонических суффиксов или не совпадать ни с одним из них. К суффиксам относится также та часть основы, где происходит чередование согласных или выпадение гласных.

Спецификация морфологии каждого входного языка содержит элементы трех типов: объекты языка (алфавит, окончания), объекты метаязыка (названия падежей, лиц, частей речи и др.) и системные числовые коды объектов метаязыка. Представление объектов метаязыка выбирается пользователем и имеет вид сокращенных названий грамматических категорий, в то время как их кодирование системными кодами фиксировано. В настоящей версии системы FEST для описания объектов метаязыка используется алфавит анализируемых текстов, в частности, кирилличный – для анализа текстов на украинском или русском языке, латинский – в случае анализа текстов, написанных с использованием латинского алфавита. Опишем кратко некоторые объекты метаязыка для спецификации грамматики русского языка.

Для спецификации лексики она делится на классы с учетом постоянных грамматических признаков лексем, а для изменяемых лексем – с учетом и окончаний. Каждый класс получает уникальное мнемоническое имя − шифр класса, который формируется из символов, обозначающих грамматические признаки. Первый символ в шифре указывает на часть речи, например: с – имя существительное, г – глагол и т.д. Следующие символы шифра определяют классификацию внутри класса, обозначенного предыдущим символом, например, для именных частей речи следующий символ шифра означает род: м – мужской, ж – женский, с – средний.

Шифр класса неизменяемых слов начинается символом '*', за которым следует указатель части речи, возможно также и указатель дальнейшей классификации в пределах указанной части речи или указатели их отношения к другим словам предложения, например, управление.

С каждым классом изменяемой лексики ассоциируется единственная конечная последовательность (кортеж) окончаний, принимаемых всеми лексемами класса. Для именных частей речи каждое окончание кортежа отвечает определенному падежу, начиная с именительного единственного числа. Для глаголов − каждое окончание отвечает определенному лицу. Число позиций во всех кортежах одинаково и равно 16, 8 позиций для единственного числа и 8 – для множественного. Позиции окончаний в кортежах для именных частей речи имеют обозначения числа и падежей, например, е-им, е-род, е-дат, е-вин, е-винж, м-род и т.д., где первая буква означает число, а следующие – названия падежей. Винительный падеж имеет две позиции – для неживых объектов и живых существ. Позиции окончаний глаголов обозначаются лицами: я, ты, он, она, оно, мы, вы, они.

Примеры шифров: сж1а-1 – шифр класса имен существительных женского рода 1–го склонения, последняя буква – окончание лексем этого класса в именительном падеже единственного числа. Другие классы существительных с такими же постоянными признаками и окончанием в именительном падеже

единственного числа имеют шифры с индексами: сж1а-2, сж1а-3, сж1а-4, потому что их кортежи в некоторых других падежах имеют различные окончания.

Классы неизменяемых частей речи не имеют кортежей.

Итак, описание морфологии любого входного языка из класса допустимых состоит из четырех разделов: алфавит, падежи и лица, шифры, кортежи окончаний.

Каждая строка первых трех разделов начинается числом – кодом раздела, за которым через пробел следует алфавит или определение метасимвола (падежа, лица или шифра). Это определение состоит из таких элементов: числовой код метасимвола, сам метасимвол с последующей точкой с запятой, далее в угловых скобках подается расширенное название (расшифровка) метасимвола.

Кортежи окончаний начинаются числами – порядковыми номерами кортежей, далее следуют шифр кортежа и сам кортеж окончаний, после которого ставится точка с запятой.

В любую строку, которая оканчивается символами '>' или ';' (точка с запятой), можно добавить комментарий, заключенный между парами символов ./* и */ без переноса на следующую строку.

Суффиксы подаются в описании лексики.

Рассмотрим несколько примеров описания объектов морфологии русского языка. Чтобы отличить элементы метаязыка от элементов самого языка, последние в этих примерах подаем курсивом.

Первый раздел – алфавит, в котором перед первой буквой стоит 0 − символ для обозначения нулевого окончания в кортежах. Строку с нулем в позиции кода раздела следует рассматривать как комментарий.

196 *0абвгдеёжзийклмнопрстуфхцчшщъыьэюя*;

197 *0АБВГДЕЁЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯ*;

Примеры описаний падежей, шифров кортежей и их кодов.

198 1 е-им; < имен. падеж ед. числа >

198 2 е-род; < род. падеж ед. числа >

198 4 е-дат; < дат. падеж ед. числа >

0 коды шифры постоянные признаки

199 1111 см1а-1ж; < имя сущ., нариц., одуш., муж. рода, 1-е склон. > /* *мужчина* */

199 1215 сж1я-5; < имя сущ., нариц., жен. рода, 1-е склон. > /* *башня* */

199 1231 сжЗь-2; < имя сущ.,нариц., жен. рода, 3 склон. > /* *функциональность* */

199 17001 *с-соед; < союз соединительный > /* *и* */

Ниже следуют примеры кортежей окончаний для слов *мужчина*, *башня, функциональность.*

0		им	рд	дт	вн	внж	тв	пр	зв	м-им	рд	дт	вн	внж	тв	пр	зв
1	см1а-1ж	*а*	*ы*	*е*	.	*у*	*ой/ою*	*е*	.	*ы*	*0*	*ам*	.	*0*	*ами*	*ах*	. ;
2	сж1я-5	*я*	*и*	*и*	*ю*	.	*ей/ею*	*и*	.	*и*	*0*	*ям*	*и*	.	*ями*	*ях*	. ;
3	сжЗь-2	*ь*	*и*	*и*	*ь*	.	*ью*	*и*	.	*и*	*ей*	*ям*	*и*	.	*ями*	*ях*	. ;

Точки в кортежах означают невозможность иметь окончание, 0 – нулевое окончание. Столбцы из трех точек относятся к звательному падежу, который не используется в научно-технических текстах.

Спецификация морфологии (файл rmorph.grm) является входной информацией для программы GENmorph, которая генерирует морфологические таблицы (файл rmorph.tbl) в таком составе:

– алфавит (нуль, малые и большие буквы);

– текстовая цепочка падежей, лиц и шифров с расшифровкой и соответствующий массив их кодов;

– цепочка окончаний, разделенных пробелами, порядковый номер первой буквы окончания – его код;

– матрица кодов кортежей и кодов окончаний, состоящая из n+1 строки и 18 столбцов, где n − число кортежей. Нулевой столбец заполнен кодами кортежей. В каждой строке с кодом кортежа размещены коды окончаний этого кортежа и адрес расшифровки шифра кортежа;

– таблица окончаний в виде деревьев, составленных из букв окончаний, начиная с конечной буквы. Деревьев столько, сколько разных конечных букв во всех окончаниях. Несколько окончаний разной длины с общим концом образуют одно дерево. Аналогичную структуру имеет и таблица суффиксов, формируемая во время построения словаря;

– таблица омонимии окончаний, в которой каждое окончание сопровождается списком кодов шифров кортежей, где присутствует окончание, и порядковыми номерами позиций этого окончания в кортежах.

Спецификация лексики

Спецификация лексики для генерации словарей представляет собой последовательность правил, оканчивающихся символом ';', за которым может следовать комментарий. Каждое правило содержит: неизменяемое слово или общую основу словоформ лексемы, или устойчивое словосочетание. Каждое слово в правиле сопровождается грамматической информацией: шифром класса, для изменяемых слов обозначением позиций в кортеже окончаний (падежи или лица), суффиксами, леммой, комментарием.

Рассмотрим простые примеры спецификации лексики русского языка (файл test.spc).

и => * : с-соед; /* союз соединительный */

для того чтобы => * : с-цел; /* союз целевой */

спис => *ок* : см2-1 "*ок*" (е-им, е-вин) "*к*";

Последнее описание следует трактовать так: в спецификации лексемы *список* в качестве основы взята неизменяемая при склонении часть слова *спис*; см2-1 − шифр кортежа окончаний лексемы – имени существительного, мужского рода, второго склонения; часть слова от основы до окончаний − суффиксы: "*ок*" в именительном и винительном падежах единственного числа и "*к*" − в остальных падежах. Чтобы получить лемму, необходимо к основе дописать суффикс *ок*.

Символ '*' в правилах для изменяемых лексем означает совпадение леммы с определяемой основой. Если лемма не совпадает с определяемой основой, то она записывается в правиле на месте символа '*'.

Прежде чем генерировать словарь по спецификации лексики визуально проверяется ее правильность по результату генерации словоформ программой GENw. Результат генерации − файл test.gen (Пример 1).

Исправленная спецификация лексем преобразуется в словарь D и структуру данных test.tbl. Словарь D − это множество словарей: test.d01, test.d02, ..., test.d20, в которых размещены цепочки символов длины соответственно 1, 2, ..., 20 со ссылками на общую структуру данных test.tbl, содержащую постоянные грамматические признаки (в текущей версии длина словарной статьи не превышает 20).

Спецификация лексики является достаточно трудоемкой задачей. В системе FEST предусмотрены средства автоматизации составления спецификаций изменяемых лексем [Мищенко, 2011]. С этой целью используются частотные списки ненайденных программой MORPH словоформ в словаре, содержащем только неизменяемую и служебную лексику. Составить спецификацию такой лексики вручную не составляет труда, и она сравнительно малочисленна. Важно и то, что выполнив эту работу однажды, такое описание можно использовать в словарях для анализа текстов разной тематики в одном и том же языке. Поэтому удобно составлять и использовать в разных целях две спецификации: неизменяемой лексики (наречия, деепричастия, предлоги, союзы) и полнозначной лексики.

Возможность создания словарей по спецификации лексики для программы MORPH особенно актуально при исследовании тематики узкоспециальных текстов ([Мищенко, 2005]).

```
D:\!!!\NADY\H\2011-PROJECTS\genw\test\genw.exe
test.gen - listing

GENW-generator
Parameters: test.spc i=rmorf-2011.tbl q=1
Time:  08:16:36      Date:  16:06:11


3  и <союз соединительный>

4  для того чтобы < союз целевой>

список (е-им, е-вин), <сущ. нариц., неодуш., муж. рода, 2 скл.>
списка (е-род), <сущ. нариц., неодуш., муж. рода, 2 скл.>
списку (е-дат), <сущ. нариц., неодуш., муж. рода, 2 скл.>
списком (е-твр), <сущ. нариц., неодуш., муж. рода, 2 скл.>
списке (е-прдл), <сущ. нариц., неодуш., муж. рода, 2 скл.>
списки (м-им, м-вин), <сущ. нариц., неодуш., муж. рода, 2 скл.>
списков (м-род), <сущ. нариц., неодуш., муж. рода, 2 скл.>
спискам (м-дат), <сущ. нариц., неодуш., муж. рода, 2 скл.>
списками (м-твр), <сущ. нариц., неодуш., муж. рода, 2 скл.>
списках (м-прдл); <сущ. нариц., неодуш., муж. рода, 2 скл.>
```

Пример 1. Компьютерный результат генерации словоформ на основе трех правил спецификации

Морфологический анализ

Морфологический анализ текстов выполняется по предложениям в два этапа.

Первый этап состоит в нахождении всех слов предложения в словарях D. Как уже отмечалось выше, МА двусторонний и начинается с поиска всего слова. В случае неудачи ищутся части слова поочередно:

сначала окончание, в случае успеха ищется оставшаяся начальная часть слова (по предположению – основа) в словаре слов меньшей длины. Если основа не найдена, осуществляется попытка найти окончание большей длины, если такового нет, ищется суффикс. Если суффикс найден, снова ищется основа меньшей длины и т. д. Результатом первого этапа МА предложения есть массив RESULT кортежей чисел, по одному кортежу для каждого слова. Кортеж массива RESULT состоит из таких данных о слове:

– адрес начала слова в тексте (порядковый номер первой буквы слова или знака препинания). Если слово не найдено, перед адресом начала слова ставится знак минус, остальные позиции кортежа заполняются нулями. То же самое выполняется и для знаков препинания;

– код шифра класса, к которому принадлежит слово;

– код падежа или лица, если окончание не имеет омонимов в кортеже, или сумма кодов падежей или лиц, если в наличии омонимия окончания;

– количество падежей или лиц: единица – если нет омонимии, и больше единицы – при наличии таковой;

– длина слова и длина основы;

– адрес леммы.

```
Частотные  <прил. относит., полная ф., муж. род>
           <им. падеж мн. числа>, LEMMA: Частотный;
списки  <сущ. нариц., неодуш., муж. рода, 2 скл.>
        <им. падеж мн. числа>, LEMMA: список;
лексем  <сущ. нариц., неодуш., жен. рода, 1 склон.>
        <род. падеж мн. числа>, LEMMA: лексема;
специальных  <прил. относит., полная ф., муж. род>
             <род. падеж мн. числа>, LEMMA: специальный;
текстов  <сущ. нариц., неодуш., муж. рода, 2 скл.>
         <род. падеж мн. числа>, LEMMA: текст;
используются  <возвр. глагол 1 спр., несов., наст.вр.>
              <3-ое лицо мн. числа>, LEMMA: использоваться;
для   <предлог>;
определения  <сущ. нариц., неодуш., сред. рода, 2 склон.>
             <род. падеж ед. числа>, LEMMA: определение;
терминологической  <прил. относит., полная ф., жен. род>
           <род. падеж ед. числа>, LEMMA: терминологический;
лексики  <сущ. нариц., неодуш., жен. рода, 1 склон.>
         <род. падеж ед. числа>, LEMMA: лексика;
.
```

Пример 2. Компьютерный результат морфологического анализа предложения

Второй этап – устранение омонимии окончаний слов предложения. Порядок расположения кортежей в массиве RESULT соответствует порядку следования слов в предложении. Это позволяет анализировать согласования соседних пар словоформ именных частей речи – в роде, числе и падеже, использовать управление с помощью предлогов и союзов, наличия переходных глаголов, знаков препинания и т. д. В подавляющем большинстве случаев омонимию окончаний удается преодолеть.

Заключительный этап – представление результатов анализа в удобном для чтения виде. После каждого найденного слова подаются постоянные признаки, затем непостоянные и лемма, если слово изменяемое.

Выше (Пример 2) подан фрагмент компьютерного результата выполнения МА одного предложения.

Заключение

Рассмотрено систему программ FEST морфологического анализа текстов на флективных и агглютинативных языках, успешно использованную в нескольких лингвистических проектах обработки научно-технических текстов на украинском и русском языках, в частности, для русско-украинского перевода и статистических исследований текстов. Проводятся эксперименты по МА текстов на украинском и русском языках с нахождением грамматических признаков слов и лемм. Систему FEST можно использовать в учебном процессе для приобретения навыков формализации знаний по морфологии любого флективного языка и введения их в компьютер для автоматического МА соответствующих текстов.

Благодарности

Автор выражает искреннюю благодарность Игорю Александровичу Мельчуку за уроки по лингвистике, полученные во время программирования его алгоритма МА для ЭВМ "Киев", а также из монографии [Мельчук, 1997] во время реализации алгоритма МА на персональных компьютерах.

Библиография

[Мельчук, 1961] Мельчук И.А. Морфологический анализ при машинном переводе (преимущественно на материале русского языка) // Сб. Проблемы кибернетики. – М.: Гос. издат. физ.-мат. лит., – 1961. – Вып. 6. – С. 207-276.

[Мищенко, 2005] Мищенко Н.М., Щеголева Н.Н. О задаче семантического индексирования тематических текстов // Proc. XI-th Intern. Conf. “Knowledge-Dialog-Solution” (June 20-30, 2005, Varna, Bulgaria). Volume II / FOI-Commerce, Sofia, 2005. P. 347-350.

[Мищенко, 2011] Мищенко Н.М., Щеголева Н.Н., Фелижанко О.Д.. Средства расширения грамматического словаря за частотным списком неизвестных слов (на украинском языке). // Мат. 5-й Межд. н/п конф. "Язык и мир: исследование и преподавание".- Кировоград, Украина. – 2011. – С. 543-547.

[Мельчук, 1997] Мельчук И.А. Курс общей морфологии. Том 1. Изд. группа "ПРОГРЕСС". Москва-Вена. 1997. 401с..

Информация об авторе

Nadiya Mishchenko – *Kyiv, Ukraine. E-mail: nadmykh@ukr.net*

Major Fields of Scientific Research: software engineering, computational linguistics