УДК 801.73:681.3

Рубашко Н.К.

РАЗРАБОТКА ЛЕКСИКО-ГРАММАТИЧЕСКИХ КЛАССИФИКАТОРОВ РУССКОГО И БЕЛОРУССКОГО ЯЗЫКОВ И ИХ ПРИМЕНЕНИЕ

Введение. В большинстве задач компьютерной лингвистики, независимо от того, статистические или символические методы используются для их решения, можно уменьшить сложность задачи, преобразовав все нечеткие лингвистические объекты в соответствии с некоторой единой процедурой в дискретные лингвистические единицы с использованием эффективной системы их кодирования и работая далее не с конкретными структурными единицами, а с их классами. Для этой цели служат специально разрабатываемые классификаторы, содержащие различные типы лингвистической информации. Неявная информация может сделаться явной путем процесса конкретной аннотации [1]. На сегодняшний день главной была и остается информация о частях речи или лексико-грамматическое кодирование, цель которого назначить каждой лексической единице код, указывающий на часть речи.

В лингвистической литературе [2, 3] подчеркивается, что доступный наблюдению язык является в форме текста, и еще никогда он не выступал ни в какой другой форме, кроме этой. Более того, как пишет Карпов В.А., «единственным источником для вывода системы=языка является генеральная совокупность текстов одного временного среза. Система, характеризующая множество текстов определенной предметной области и выведенная из этого множества как из лингвистического универсума, будет порождать и сам универсум, и процесс его порождения, т.е. речевую деятельность» [2, с.25].

В исследованиях естественного языка (ЕЯ) и при разработке систем автоматической обработки текста (АОТ) важно иметь аннотированный, или базовый, словарь, в котором каждому слову назначены вне контекста все допустимые лексико-грамматические коды (ЛГК), и аннотированный корпус текстов (КТ), в котором каждому слову приписан ЛГК, единственно допустимый в данном контексте [4], что делает его не текстом, где лингвистическая информация неявно представлена, а текстом, который может рассматриваться как хранилище лингвистической информации [1].

Базовый словарь является основой лингвистического анализа, а корпус текстов используется для получения количественных измерений ЕЯ, тестирования лингвистических гипотез и систем АОТ.

Принципы разработки лексико-грамматического классификатора. Для ЕЯ различают такие структурные единицы, как морфема, словоформа, конфигурация, фраза, предложение, дискурс, текст [5]. Каждая из этих структурных единиц образуется на основе определенных правил заданного ЕЯ. В соответствии с этими правилами каждой j-ой структурной единице ЕЯ i-го уровня, обозначим $t_j^{(i)}$, может быть поставлено в соответствие множество $S_j^{(i)}$ ее морфологических, синтаксических и семантических свойств, известное под названием *кода* (*тэга* (англ. *tag*), *марки, метки)*. Такой процесс в литературе называют *аннотированием*, *маркированием*, *разметкой*, *тегированием* словаря (текста) ЕЯ, а множество кодов (*tagset*) – *классификатором* [1, 6, 7, 8].

В зависимости от поставленной задачи структурные единицы ЕЯ любого из указанных уровней с соответствующими множествами свойств могут быть взяты в качестве элементов словаря $(t_j^{(i)}, S_j^{(i)})$ (в самом общем смысле этого слова), содержащего данные о грамматике языка, характеризующие категориальный состав грамматических форм и конструкций, и свод правил, определяющих условия применимости в определенном структурном или лексико-семантическом контексте. Для определения подходящего множества кодов требуется значительная лингвистическая экспертиза, и создание словаря с приписанными его элементам кодами – очень сложная задача.

В настоящее время не существует стандартов на представление подобной информации, однако на основе проведенного анализа можно сделать следующие предположения относительно системы кодов:

а) код должен быть экономным и в то же время избыточным;

б) код должен снимать омонимию, т.е. быть однозначным;

в) код должен обеспечивать:

- наличие нескольких “слоев” информации, извлекаемых из разметки независимо друг от друга;
- потенциальную расширяемость на типы информации, не охватываемые аннотацией на определенном этапе.

На уровне слова используется основной тип лингвистического кодирования – морфологический анализ или аннотирование по частям речи (*part-of-speech annotation*). Он направлен на то, чтобы с минимальными потерями информации получить достоверное представление текстов различных предметных областей на уровне отдельных слов [7]. Подобный вид кодирования увеличивает определенность поиска данных в корпусах текстов и формирует основу для синтаксического и семантического анализа.

Данный вид кодирования был одним из первых типов аннотирования, примененных в компьютерной лингвистике, и остается на сегодня главным, поскольку существуют компьютерные программы, позволяющие выполнять автоматическое маркирование слов по частям речи с высокой степенью точности [1].

Тем не менее, даже для одного языка может существовать несколько разных систем кодирования по частям речи. Все зависит от конкретных приложений, т.е. от того, какая степень детализации требуется, и от природы обрабатываемого языка. Например, для английского языка в British National Corpus количество кодов колеблется от 58 (C5 tagset) до 138 (C6 tagset) [6], а базовый классификатор словенского языка, являющегося флективным языком, содержит около 1600 кодов [9].

При разработке классификатора для белорусского и русского языков были учтены принципы кодирования, изложенные в [7], но применительно к языкам с разветвленной флективной системой. Оба языка имеют универсально алгоритмичный для данной пары языков классификатор благодаря общности морфологических категорий.

При создании классификатора использовалось подразделение слов на лексико-грамматические классы, называемые традиционно частями речи: имя существительное, имя прилагательное, глагол и др. [10, 11], исходя из того, что если набор грамматических признаков, описывающих слово и составляющих его характеристику, представить в виде иерархической структуры, то высший ярус ее займет признак части речи, поскольку он покрывает практически всю лексику [12].

Далее при разработке классификатора учитывались не только классы слов, но и подклассы. Например, местоимения распадаются на ряд подклассов, различных по лексическим формам и синтаксическим функциям (личные местоимения, возвратные местоимения, притяжательные местоимения и т.д.).

С точки зрения синтаксического функционирования определенных классов слов были выделены устойчивые словосочетания, рассматриваемые как единое целое. Например, предлоги *в соответствии с* (рус.), *нягледзячы на* (бел.).

Полученный код содержит от двух до шести символов, образующих группы с частично иерархической структурой. В состав кода входит основа кода, суффиксы, обозначающие лексический подкласс, и суффиксы грамматических подклассов. Полный код слова образуется по правилу:

<значение кода>::= <основа кода> [<суффикс лексического подкласса>…] [<суффикс грамматического подкласса>…]

Основы кодов и их значения приведены в табл. 1, суффиксы лексических подклассов и их значения – в табл. 2, суффиксы грамматических подклассов и их значения – в табл. 3.

***Рубашко Наталья Константиновна**, ассистент кафедры математического обеспечения АСУ факультета прикладной математики и информатики Белорусского государственного университета.*

Беларусь, БГУ, 220050, г. Минск, пр. Независимости, 4.

Таблица 1. *Основы кодов частей речи (русский и белорусский языки)*

Основа кода	Значение кода	
CC	сочинительный союз (Conjunction Coordinating)	Служебные части речи
CS	подчинительный союз (Conjunction Subordinating)	
CM	сравнительный союз (Conjunction coMparative)	
IN	предлог (prepositIon)	
IR	предлог – омограф наречия	
IV	предлог – омограф деепричастия	
TC	частица (parTiCle)	
TP	постпозиционная частица (parTicle Postpositive)	
XNOT	отрицательная частица	
AB…	сокращенное наименование (ABbreviation)	
MD	модальное слово (MoDal word)	
UH	междометие (interjection)	
WD	вводное слово (parenthetic WorD)	
B…	деепричастие (verBal adverb)	
D…	числительное (numeral)	
J…	имя прилагательное (adJective)	
L…	причастие (participLe)	
N…	имя существительное (Noun)	
P…	местоимение (Pronoun)	
R…	наречие (adveRb)	
V…	глагол (Verb)	

Таблица 2. *Суффиксы лексических подклассов*

Суффикс	Значение	Употребление
C	количественное (Cardinal)	с числительным
L	порядковое (coLlective)	
O	собирательное (Ordinal)	
Q	качества, образа действия (Quality)	с наречием
M	количества, меры и степени (Measure)	
T	времени (Time)	
L	места, направления (Locality)	
V	причины (motiVe)	
P	цели (Purpose)	
R	сравнительная степень (compaRative)	с наречием, прилагательным
T	превосходная степень (superlaTive)	
J	качественное (qualitative)	
R	относительное (Relative)	
H	краткая форма (sHort)	
$	притяжательное (possessive)	с прилагательным, местоимением
P	совершенный вид (Perfective aspect)	с глаголом
I	несовершенный д (Imperfective aspect)	
L	возвратное (refLexive)	с глаголом, причастием, местоимением
A	действительное (Active)	с причастием
P	страдательное (Passive)	
Суффикс	**Значение**	**Употребление**
N	нарицательное (commoN)	с существительным
P	собственное (Proper)	
A	одушевленное (Animate)	с существительным
I	неодушевленное (Inanimate)	
P	личное (Personal)	с местоимением
D	указательное (Demonstrative)	
Q	вопросительное (Questioning)	
C	относительное (Comparative)	
T	определительное (deTerminative)	
N	отрицательное (Negative)	
R	неопределенное (indeteRminate)	
S	единственное число только (Sole)	с местоимением

Полученный классификатор представляет собой систему лексико-грамматических свойств белорусского и русского языков и имеет два уровня. Первый уровень включает 104 кода и характеризует словоизменительную парадигму в целом, т.е. это код, который одинаков у всех слов из парадигмы. Второй уровень содержит 105 кодов и характеризует каждое слово в парадигме – словоформу, так как содержит грамматическую информацию.

Совокупность кодов первого и второго уровней и образует уникальный код, который приписывается конкретной словоформе (всего таких кодов 1007).

Таблица 3. *Суффиксы грамматических подклассов*

Суффикс	Значение	Употребление
M	мужской род (Masculine)	со всеми частями речи, кроме служебных
F	женский род (Feminine)	
N	средний род (Neuter)	
P	множественное число (Plural)	
O	именительный падеж (nOminative)	с числительным, существительным, прилагательным, местоимением, причастием
G	родительный падеж (Genitive)	
D	дательный падеж (Dative)	
A	винительный падеж (Accusative), неодушевл.	
U	винительный падеж (accUsative), одушевл.	
I	творительный падеж (Instrumental)	
R	предложный падеж (pRepositional)	
1	1 лицо	с глаголом, местоимением
2	2 лицо	
3	3 лицо	
R	настоящее время (pResent)	с глаголом, причастием
P	прошедшее время (Past)	
F	будущее время (Future)	
I	изъявительное наклонение (Indicative mode)	*с глаголом*
M	повелительное наклонение (iMperative mode)	

Для служебных частей речи при формировании кода используется только основа кода. Коды частей речи со словоизменительной парадигмой образуются по следующим правилам:
а) для существительного, местоимения – часть речи, категория, [одушевленность / неодушевленность], род, число, падеж;
б) для прилагательного, причастия, некоторых местоимений, порядкового числительного – часть речи, категория, род, число, падеж;
в) для глагола – часть речи, вид, [возвратность], наклонение, время, [лицо], число, [род].

Например, NNAMO – имя существительное, нарицательное, одушевленное, мужского рода, ед. ч., в именительном падеже; VPLIR1 – возвратный глагол совершенного вида, изъявительного наклонения, настоящего времени, 1-го лица, ед.ч.

Предложенное кодирование позволяет однозначно закодировать местоимения, прилагательные, существительные, причастия и глаголы так, что все виды омонимии (грамматическая, межклассовая и внутриклассовая, внутрипарадигматическая и межпарадигматическая) внутри одного класса снимаются кодом, особенно в случаях согласования в роде, в роде и числе, в роде, числе и падеже.

Помимо описанного классификатора разработана система специальных кодов, предназначенных для маркирования знаков препинания, служебной информации. Знаки препинания для облегчения обработки кодируются символом, совпадающим с самим знаком.

Применение лексико-грамматического классификатора для аннотирования словаря и корпуса текстов. На основе разработанного классификатора было выполнено кодирование русского и белорусского машинных словарей словоформ и аннотирование корпусов текстов для двух указанных языков, являющихся государственными языками Республики Беларусь.

Указанный классификатор, словари и корпусы текстов легли в основу Компьютерного фонда белорусского языка, разрабатывавшегося в рамках Государственной программы информатизации «Электронная Беларусь».

Поскольку, как отмечалось выше, для решения задач компьютерной лингвистики нужны и базовый словарь, и корпус текстов, то, прежде всего, необходимо определиться с принципиальной схемой выполнения аннотирования.

Сначала разрабатывается классификатор свойств $K^{(r_0)}$ языка L, где r_0 – начальный уровень (чаще всего слов). Затем выполняется разработка и аннотирование базового словаря D, где каждой словоформе приписывается множество соответствующих ему (вне текста) свойств (кодов) из $K^{(r_0)}$; полученный словарь обозначим $D^{(r_0)}$. Используя словарь $D^{(r_0)}$, осуществляется автоматическая идентификация исходного корпуса текстов T_0: каждому словоупотреблению из T_0 приписывается соответствующее ему по словарю $D^{(r_0)}$ множество кодов.

На заключительном этапе (вручную или автоматически с последующей корректировкой) снимается полученная многозначность кодов – каждому словоупотреблению соответствует единственный допустимый в данном контексте код.

В итоге формируется аннотированный корпус текстов $T_L^{(r_0)}$ ЕЯ L уровня r_0, в котором для каждого словоупотребления текста указан код класса в соответствии с классификатором $K^{(r_0)}$.

Для выполнения аннотирования корпусов текстов был разработан специальный инструментарий – АРМ разработчика корпуса текстов на уровне r_0, который работает в соответствии со схемой, описанной выше. В качестве $K^{(r_0)}$ используется классификатор лексико-грамматических свойств. Кроме классификатора используются либо белорусский, либо русский словарь D. Выбор языка обработки осуществляется на этапе запуска процесса аннотирования. Использование готового словаря делает ненужным формирование словаря разработчиком КТ.

Таким образом, были разработаны базовые словари для белорусского и русского языков.

Общие размеры словарей составили: для белорусского языка – около 2,7 млн. словоупотреблений, для русского – около 4 млн. Расчет производился для словоупотреблений, а не для слов, так как указанные языки принадлежат к флективным языкам и характеризуются богатой словоизменительной парадигмой.

Далее, в соответствии с указанной принципиальной схемой были разработаны аннотированные корпусы текстов для белорусского и русского языков.

Общие размеры корпусов составили: для белорусского языка – около 400 тыс. словоупотреблений, для русского – около 1 млн. Расчет производился также для словоупотреблений. Для расчета среднего количества кодов для каждого словоупотребления из рассмотрения были исключены знаки препинания, формулы и иностранные слова.

На основе полученных словарей и корпусов текстов было проведено исследование распределения кодов для указанных языков.

Распределение кодов для белорусского языка приведено в табл. 4, для русского языка – в табл. 5.

Полученные результаты показывают, что среднее значение количества кодов для одного словоизменения одинаково как для русского, так и белорусского языков, что объясняется тем фактом, что указанные языки принадлежат к группе родственных языков. И это несмотря на наличие для обоих языков словоупотреблений, имеющих количество кодов, больше 10.

Таблица 4. *Распределение кодов для белорусского языка*

Базовый словарь		Корпус текстов	
Количество кодов у слова	**Количество слов**	**Количество кодов у слова**	**Количество слов**
1	804109	1	38116
2	311371	2	5840
3	115608	3	1062
4	51835	4	294
5	46286	5	99
6	5374	6	40
7	444	7	15
8	1687	8	6
9	159	9	2
10	1501	10	1
11	19	14	1
12	582	28	1
13	25	33	1
14	14	34	1
15	8		
16	1	Попарно различных словоупотреблений	45478
17	2		
18	7	Словоупотреблений с различными кодами	55183
20	2		
24	17	Среднее количество кодов на словоупотребление	**1,2**
27	6		
28	2		
29	1		
30	2		
31	1		
36	3		
38	1		
39	1		
Попарно различных словоупотреблений	1339067		
Словоупотреблений с различными кодами	2286715		
Среднее количество кодов на словоупотребление	**1,7**		

Таблица 5. *Распределение кодов для русского языка*

Базовый словарь		Корпус текстов	
Количество кодов у слова	**Количество слов**	**Количество кодов у слова**	**Количество слов**
1	1153308	1	78365
2	493162	2	11423
3	246367	3	2076
4	83179	4	627
5	8135	5	203
6	6592	6	73
7	181	7	33
8	1209	8	13
9	87	9	4
10	23	10	5
11	7	11	2
12	640	12	1
13	17	13	3
14	10	14	1
15	2	15	3
16	1	17	1
18	12	31	1
24	17	33	1
25	1		
26	174	Попарно различных словоупотреблений	92834
27	198		
28	52	Словоупотреблений с различными кодами	112034
29	5		
30	6	Среднее количество кодов на словоупотребление	**1,2**
31	1		
36	3		
39	1		
63	1		
Попарно различных словоупотреблений	1993390		
Словоупотреблений с различными кодами	3324333		
Среднее количество кодов на словоупотребление	**1,7**		

Заключение. Разработанные классификатор, словари и корпусы текстов могут быть использованы в различных задачах автоматической обработки ЕЯ: корректировки орфографии, машинного перевода, автоматического реферирования, информационного поиска и др., на начальном этапе лингвистического анализа текста (запроса пользователя) с целью аннотирования текста лексико-грамматическими кодами.

Кроме того, лексико-грамматические коды позволяют оптимизировать разработку так называемых патернов при выполнении синтаксического анализа текстов, поскольку обеспечивают их обобщение на уровне ЛГК, основанное преимущественно на морфологических признаках словоформ и правилах их комбинации.

СПИСОК ЦИТИРОВАННЫХ ИСТОЧНИКОВ

1. Mcenery T., Wilson A. Corpus Linguistics. – Edinburgh: Edinburgh University Press, 1996. – 132 p.
2. Карпов В.А. Язык как система. – Мн.: Вышэйшая школа, 1992. – 302 с.
3. Лаптева О.А. Дискретность в устном монологическом тексте// Русский язык: Текст как целое и компонента текста (Виноградовские чтения XI). – М.: Наука, 1982. – С.77.
4. Богуславский И.М., Григорьев Н.В., Григорьева С.А. и др. Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации // Труды Международного семинара Диалог'2000 по компьютерной лингвистике и ее приложениям. – Протвино, 2000. – Том 2. – С. 41–47.
5. Методы автоматического анализа и синтеза текста / Пиотровский Р.Г., Билан В.Н., Боркун М.Н., Бобков А.К. – Мн.: Вышэйшая школа, 1985. – 222 с.
6. Leech G., Garside R., Bryant M. CLAWS4: The tagging of the British National Corpus // Proceedings of the 15th International Conference on Computational Linguistics (COLING 94) Kyoto, Japan, 1994. – P.622-628.
7. Совпель И.В. Инженерно-лингвистические принципы, методы и алгоритмы автоматической переработки текста. – Мн.: Вышэйшая школа, 1991. – 118 с.
8. Богуславский И.М., Григорьев Н.В., Григорьева С.А. и др. Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации // Труды Международного семинара Диалог'2000 по компьютерной лингвистике и ее приложениям. – Протвино, 2000. – Том 2. – С. 41–47.
9. Vintar Š. A Lexical Analysis of the IJS-ELAN Slovene-English Parallel Corpus // Vintar, Š. (ed.) Proceedings of the workshop Language Technologies –Multilingual Aspects. Ljubljana: Faculty of Arts., 1999.
10. Курс беларускай мовы: Падручник / Л.І. Сямешка, І.Р. Шкраба, З.І. Бадзевіч. – Мн.: Універсітэцкае, 1996. – 654 с.
11. Русский язык. Часть I. Изд. 3-е, испр. и доп. / А.М. Бордович, Н.И. Гурский, Е.С.Хмелевская, Э.К. Бирилло. – Мн.: Вышэйшая школа, 1977. – 416 с.
12. Григорян В.М. Соотношение категорий субстантивности и залоговости // Русский язык: Проблемы грамматической семантики и оценочные факторы в языке (Виноградовские чтения XIX–XX). – М.: Наука, 1991. – С.3–11.

Материал поступил в редакцию 22.01.08

RUBASHKO N.K. Development of lexical and grammatical qualifiers of Russian and Byelorussian languages and their application

The purpose of the present job is the description of principles of development of lexical and grammatical qualifiers for Russian and Byelorussian languages and their use at creation of the dictionaries and cases of the texts.

УДК 004.89:004.4

Постаногов Д.Ю.

РАСПОЗНАЮЩИЕ ШАБЛОНЫ НА ОСНОВЕ РАСШИРЕННЫХ РЕГУЛЯРНЫХ ВЫРАЖЕНИЙ В ЗАДАЧАХ АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТА

Введение. Эффективность применения интеллектуальных информационных систем (ИИС) главным образом определяется качеством и количеством знаний, содержащихся в используемых ими базах знаний (БЗ) [1] [2] [3]. Возможность разработки действенной БЗ достаточно большого объема, в свою очередь, обуславливается способами организации взаимодействия разработчиков ИИС, как правило – инженеров по знаниям, экспертов и программистов [4].

Одним из базовых компонентов ИИС и, прежде всего, различных систем автоматической обработки текстов является лингвистический процессор (ЛП), основу которого, в свою очередь, составляет лингвистическая база знаний (ЛБЗ). Важным аспектом в обеспечении эффективности их разработки и оптимального устройства является степень интегрированности лингвистических знаний о естественном языке (ЕЯ) в программный код системы. Данный вопрос касается как *декларативной* части ЛБЗ, к которой можно отнести словари языка и списки различного назначения, вероятностные модели и другие данные статического характера, так и наиболее значительной, *процедурной* части ЛБЗ, которую составляют правила различного рода в зависимости от соответствующего им этапа анализа текста и структурного уровня ЕЯ, например, правила лексико-грамматического, синтаксического и семантического анализа предложений, в совокупности и в сочетании с декларативной частью ЛБЗ представляющие собой определенную лингвистическую модель ЕЯ.

Очевидно, что программист, описывающий алгоритм работы ЛП на определенном языке программирования, чаще всего, не обладая глубокими специальными знаниями о структуре ЕЯ, не способен самостоятельно в сжатые сроки реализовать достаточное количество лингвистических правил, которые бы обеспечили обработку текста с высоким качеством. Более того, организация взаимодействия программиста с экспертом–лингвистом, который лишь передает ему свои знания о ЕЯ напрямую либо через посредничество инженера по знаниям, также не позволяет достичь высокой эффективности разработки ЛП в силу необходимости неоднократного внесения правок в программный код при выполнении трудоемких процедур проверки используемых лингвистических гипотез, тестирования и отладки системы. В этой связи наиболее перспективным подходом к разработке процедурной части ЛБЗ является явное обособление лингвистических правил от программного кода за счет использования проблемно-ориентированных языков [5] как средств описания знаний о ЕЯ самим экспертом-лингвистом. Данный подход обеспечивает открытость и расширяемость системы в смысле возможности явного разделения труда экспертов данной предметной области, т. е. лингвистов, которые описывают декларативные и процедурные знания о ЕЯ на таких языках, и программистов, разрабатывающих аппарат их интерпретации [6]. Основными задачами инженера по знаниям в этом случае являются: взаимодействие с экспертом-лингвистом с целью определения требований к описательным возможностям этих языков, проектирование соответствующих систем обозначений — *нотаций*, а также описание способов их интерпретации, в дальнейшем реализуемых программистом в виде соответствующих алгоритмов.

В данной статье рассматриваются вопросы разработки проблемно-ориентированных языков и аппарата их эффективной интерпретации для описания систем так называемых *шаблонов*, являющихся одним из формальных представлений лингвистических правил и имеющих чрезвычайно широкий спектр применения в теории и практике инженерно-лингвистического подхода [7] к обработке ЕЯ.

Применение регулярных выражений в шаблонах. В общем случае правила ЛБЗ реализуются в виде так называемых распознающих шаблонов, которые главным образом сводятся к форме

***Постаногов Денис Юрьевич**, начальник отдела разработки средств интеллектуализации информационных систем ИП «Инвеншион Машин».*