УДК 62.506.2

М. Ф. БОНДАРЕНКО, канд. техн. наук,
А. Ф. ОСЫКА, инж.

АЛГОРИТМ МОРФОЛОГИЧЕСКОГО АНАЛИЗА ПОРЯДКОВЫХ ЧИСЛИТЕЛЬНЫХ РУССКОГО ЯЗЫКА

Целью данной работы является моделирование психической деятельности человека при выделении из последовательности словоформ русского языка в письменной речи простых и сложных порядковых числительных и при установлении их характеристик. Предлагаемая ниже модель справедлива для всего множества словоформ русского языка, причем эти словоформы могут быть образованы от слов, помещенных в какой-либо словарь, и от «псевдослов».

Несколько слов о понятии «порядковые числительные». В лингвистике отсутствует единое мнение о том, относить ли слова типа «пятый», «семидесятый», «двадцатипятимиллионный» к порядковым числительным или к относительным прилагательным [1, 2]. Все зависит от признаков, которые положены в основу той или иной классификации слов по частям речи.

При построении алгоритмов автоматического распознавания частей речи по формальным признакам было решено отнести слова данного типа к числительным, так как в качестве формальных признаков числительных принимались не только окончания, но и основы, число которых ограничено. Такой выбор формальных признаков позволяет: а) однозначно выделить все числительные, в том числе порядковые; б) при определении числительных и установлении их характеристик избегать присвоения входному слову какой-либо сопроводительной информации; в) основы числительных рассматривать как носители элементарных единиц смысла, что дает необходимую информацию для семантического анализа («понимания» значения числительного); г) информацию об основах и окончаниях числительных, хранимую в памяти ЭВМ для анализа числительных, применять в целях их синтеза.

В процессе исследования порядковых числительных были выделены их основы (табл 1.) Они разбиты на 13 групп-списков. Основа числительного включалась в тот или иной список на основании следующих признаков: 1) обозначение одних и тех же разрядов чисел (единицы, десятки, сотни, тысячи и больше); 2) сочетаемость с одними и теми же основами из других списков. В некоторые списки вошло только по одной основе из-за особенностей сочетаемости этих основ. В графе j табл. 1 помещены номера основ ($j=1, \ldots, 43$), в графе Z_j — основы порядковых числительных, в r_j — длины в буквах j-х основ, в графе θ_{jt} — — номера типов склонений для j-х основ ($j=1, \ldots, 22$; $t=1, \ldots, 6$), которые могут стоять в простых порядковых числительных или в конце сложного порядкового числительного. Для

основ, которые всегда стоят в начале или в середине сложных порядковых числительных, номера типов склонения не приводятся.

Таблица 1

j	Z_j	r_j	Θ_{jt}					
1	2	3	4					
1	*девяност*	8	4	6	7	8	17	
2	*сороков*	7	1	5	6	7	8	
3	*четверт*	7	4	6	7	8	17	
4	*восьм*	5	1	5	6	7	8	
5	*девят*	5	4	6	7	8	17	
6	*десят*	5	4	6	7	8	17	
7	*седьм*	5	1	5	6	7	8	
8	*втор*	4	1	5	6	7	8	
9	*перв*	4	4	6	7	8	17	
10	*трет*	4	2	9	10	11	12	
11	*шест*	4	1	5	6	7	8	
12	*пят*	3	4	6		8	17	
13	*сот*	3	4	6	7	8	17	
14	*сот*	3	4	6	7	8	17	18
15	*дцат*	4	4	6	7	8	17	19
16	*десят*	5	4	6	7	8	17	19
17	*надцат*	7	4	6	7	8	17	19
18	*миллиард*	8	3	13	14	15	16	18
19	*триллион*	8	3	13	14	15	16	18
20	*биллион*	7	3	13	14	15	16	18
21	*миллион*	7	3	13	14	15	16	18
22	*тысяч*	5	3	13	14	15	16	20
23	*восем*	5				—		
24	*девят*	5				—		
25	*четыр*	5				—		
26	*шест*	4				—		
27	*две*	3				—		
28	*пят*	3				—		
29	*сем*	3				—		
30	*три*	3				—		
31	*два*	3				—		
32	*четырех*	7				—		
33	*девяти*	6				—		
34	*восьми*	6				—		
35	*шести*	5				—		
36	*пяти*	4				—		
37	*семи*	4				—		
38	*двух*	4				—		
39	*трех*	4				—		
40	*одно*	4				—		
41	*сто*	3				—		
42	*девяносто*	9				—		
43	*сорока*	6				—		

Проанализируем значение термина «основа порядкового числительного» в данной работе. В список 8 (табл. 1) помещены основы *двух, трех, четырех* и т. д. Они представляют собой фор-

му родительного падежа простых количественных числительных, которые можно было бы членить и дальше. Но целью нашего исследования является модель распознавания порядковых числительных в письменной речи. Эта модель положена в основу программы для ЭВМ, в связи с чем немаловажное значение имеет объем машинной памяти и машинного времени.

В письменной речи в сложных порядковых числительных при обозначении количества сотен или единиц всегда употребляется *двух, трех, четырех* и т. д. Дальнейшее членение этих составных частей порядковых числительных в процессе помещения их в словари может значительно усложнить алгоритм распознавания порядковых числительных и увеличить объем машинной памяти и машинного времени. При этом результат на выходе будет тем же.

Таблица 2

m	γ_m	φ_m
1	1	13
2	14	14
3	15	15
4	16	16
5	17	17
6	18	22
7	23	30
8	32	39
9	40	40
10	41	41
11	30	31
12	34	37
13	42	43

В табл. 1 основы Z_j сгруппированы по спискам. Информация о том, в какой список входит данная основа, содержится в табл. 2. В графе m помещены номера списков ($m = 1, ..., 13$), в графе γ_m — номера строк j (табл. 1), с которых начинаются списки m, в графе φ_m — номера строк (табл. 1), которые являются последними в списках m. Некоторые основы из табл. 1 входят в два списка. Это сделано для сокращения объема хранимой информации.

При исследовании окончаний порядковых числительных было выделено 17 типов их склонения. Окончания выделенных типов склонения порядковых числительных приведены в табл. 3. Окончания 18-го, 19-го, 20-го типов склонения в табл. 3 принадлежат количественным числительным. Они используются для проверки правильности падежных форм основ *-сот-, -десят-, -надцат-, -дцат-*, когда эти основы стоят в середине слова. В графе S табл. 3 помещены номера типов склонений ($S = 1, ..., 20$). В графе O_{sv} ($S = 1, ..., 20$; $v = 1, ..., 6$) даются окончания именительного, родительного, дательного, винительного, творительного и предложного падежей S-х типов склонения числительных. В графе J_s представлены некоторые сведения о существительных, с которыми сочетается порядковое числительное, если оно изменяется по S-му типу склонения.

На вход алгоритма могут подаваться словоформы не только слов, помещенных в какой-либо словарь, но и псевдослов, например, *двухсемидесятисотый*. Поэтому простое вычленение и

распознавание правильных основ и окончаний порядковых числительных еще не свидетельствует о том, что анализируемое слово является порядковым числительным. В связи с этим для правильной идентификации порядковых числительных необходимо иметь набор составных элементов (основ и окончаний) и описание допустимого порядка их следования в слове. Такое описание в виде графа показано на рис. 1.

Таблица 3

S	O_{sv}						J_s
	Им.	Род.	Дат.	Вин.	Твор.	Пред.	
1	*ой*	*ого*	*ому*	*ого*	*ым*	*ом*	м. р. одуш.
2	*ий*	*ьего*	*ьему*	*ьего*	*ьим*	*ьем*	м. р. одуш.
3	*ный*	*ного*	*ному*	*ного*	*ным*	*ном*	м. р. одуш.
4	*ый*	*ого*	*ому*	*ого*	*ым*	*ом*	м. р. одуш.
5	*ой*	*ого*	*ому*	*ой*	*ым*	*ом*	м. р. неодуш.
6	*ая*	*ой*	*ой*	*ую*	*ой*	*ой*	ж. р. неодуш.
7	*ое*	*ого*	*ому*	*ое*	*ым*	*ом*	ср. р. неодуш.
8	*ые*	*ых*	*ым*	*ые*	*ыми*	*ых*	pl. t. неодуш.
9	*ий*	*ьего*	*ьему*	*ий*	*ьим*	*ьем*	м. р. неодуш.
10	*ья*	*ьей*	*ьей*	*ью*	*ьей*	*ьей*	ж. р. неодуш.
11	*ье*	*ьего*	*ьему*	*ье*	*ьим*	*ьем*	ср. р. неодуш.
12	*ие*	*ьих*	*ьим*	*ие*	*ьими*	*ьих*	pl. t. неодуш.
13	*ный*	*ного*	*ному*	*ный*	*ным*	*ном*	м. р. неодуш.
14	*ная*	*ной*	*ной*	*ную*	*ной*	*ной*	ж. р. неодуш.
15	*ное*	*ного*	*ному*	*ное*	*ным*	*ном*	ср. р. неодуш.
16	*ные*	*ных*	*ным*	*ные*	*ными*	*ных*	pl. t. неодуш.
17	*ный*	*ного*	*ному*	*ый*	*ым*	*ом*	м. р. неодуш.
18	#	#	—	#	—	—	—
19	*ь*	*и*	*и*	*ь*	*ью*	*и*	—
20	*а*	*и*	*е*	*у*	*ей*	*е*	—

При построении графа, описывающего структуру любого порядкового числительного, которое записывается (или может быть записано) одним словом, мы исходили из следующих условий: 1) сложные и составные порядковые числительные пишутся аналогично соответствующим количественным числительным: *семнадцатый, семьдесят девятый* [3]; 2) порядковые числительные, оканчивающиеся на *-сотый, -тысячный, -миллионный* и т. п., пишутся слитно [4]; 3) порядковые числительные, имеющие семь и более основ, не принято обозначать буквами. В таких порядковых числительных пишется лишь последний элемент [5], например: 1375-*миллионный*. Узел данного графа представляет собой основу, входящую в список, номер которого стоит в узле. Так, ветвь графа на рис. 2 моделирует 1280 различных шести основных порядковых числительных, которые в тексте могут быть представлены 14 080 различными словоформами, например: *двухсотдвадцатипятимиллионный*.

Данный граф, описывающий порядок следования основ в простых и сложных порядковых числительных, был положен в основу блок-схемы алгоритма определения однословных порядковых числительных русского языка в письменной речи (см. рис. 1). Слово X,

поступающее на вход алгоритма, рассматривается как последовательность длиной n из элементов $a_к$:

$$X = a_1,\ a_2,\ \ldots,\ a_n.$$

Под $a_к$ понимается любая буква русского алфавита, стоящая на $к$-м месте в слове X.

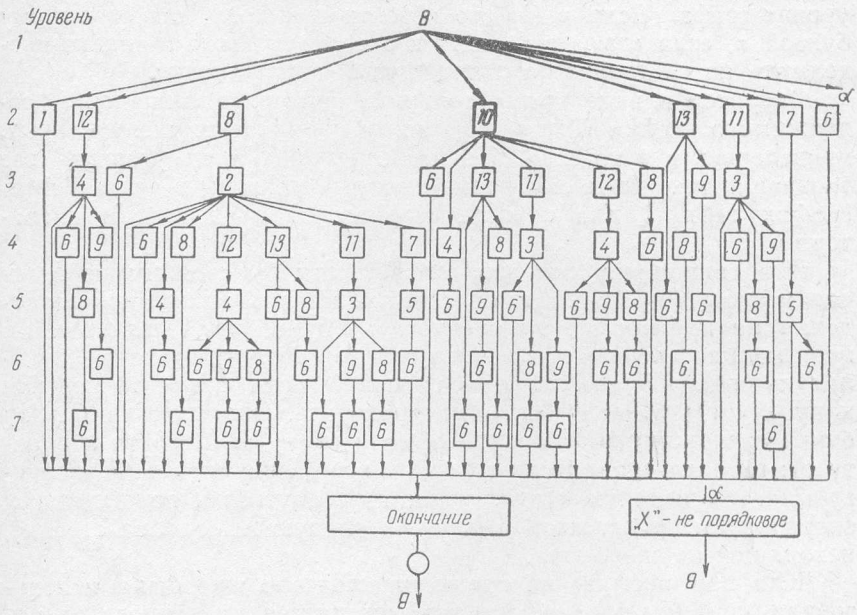

Рис. 1.

Блок-схема алгоритма представляет собой граф, узлы которого расположены на девяти уровнях и представляют собой операторы пяти видов. Оператор ввода B записывает исходное слово X в рабочую ячейку W. Анализ начинается по крайней слева, направленной вниз стрелке; Я — оператор выхода. Алгоритм имеет два

Рис. 2.

выхода: один для слов — порядковых числительных, второй — для слов, которые не являются порядковыми числительными. Оператор в виде квадрата проверяет совпадение первой части содержимого рабочей ячейки W с одной из основ числительных, помещенных в списке, номер которого стоит в середине квадрата. В случае совпадения из содержимого W удаляется выделенная основа числительного, и анализ продолжается по крайней слева, направленной вниз из данного оператора стрелке. В случае несовпадения

первой части содержимого W ни с одной из основ в списке следует возвратиться к предыдущему оператору и продолжать анализ по следующей справа стрелке, исходящей из того же оператора. Если при анализе на предыдущем этапе уже была использована крайняя справа стрелка, исходящая из данного оператора, то необходимо проверить, находится ли данный оператор на втором уровне графа. Если «нет», то перейти к выходу, обозначенному буквой α, если «да», то перейти к оператору ввода и анализ продолжать по следующей справа, идущей вниз стрелке.

Оператор в виде прямоугольника проверяет совпадение содержимого ячейки W с окончаниями порядковых числительных, приведенными в табл. 3. Такому сравнению подвергаются все окончания тех типов склонения, которые указаны в первых пяти графах табл. 1 для основы, выделенной предыдущим оператором.

При совпадении содержимого W хотя бы с одним из окончаний анализ продолжать по стрелке, идущей вниз. Если совпадения не произошло, то следует: а) проверить, является ли предыдущая выделенная основа основой *десят* из списка *4*, *надцат* из списка 5 или *дцат* из списка 3; если «нет», то перейти к пункту г); б) если «да», то проверить, является ли первая буква содержимого W буквой *и*; если «нет», то перейти к пункту г); в) если «да», то из содержимого W удалить первую букву *и*, затем перейти к пункту г); г) вернуться к предыдущему оператору и продолжать анализ по следующей справа стрелке, исходящей из оператора.

Если при анализе на предыдущем этапе уже была использована крайняя справа стрелка, исходящая из данного оператора, то необходимо проверить, находится ли данный оператор на втором уровне графа. Если «нет», то перейти к выходу, обозначенному буквой α. Если «да», то перейти к оператору ввода и анализ продолжать по следующей справа идущей вниз стрелке.

Оператор в виде кружка формирует информацию об анализируемом порядковом числительном. Информация о падеже выдается в виде цепочек по шесть знаков, состоящих из нулей и единиц. Одна цепочка соответствует окончаниям (см. табл. 3) какого-то одного типа склонения S порядковых числительных, номер которого указан в табл. 1 для последней выделенной основы. Нуль ставится в соответствие окончанию, которое проверялось на совпадение с содержимым W и не совпало. Единица ставится в соответствие окончанию, которое совпало с содержимым W. Цепочки, состоящие из одних нулей, исключаются. Цепочки, соответствующие окончаниям S-го типа склонения, дополняются данными строки S, графы J_s табл. 3, т. е. сведениями о роде и числе, и сведениями об одушевленности и неодушевленности существительных, с которыми может сочетаться данное числительное.

По данной блок-схеме алгоритма составлена программа для ЭВМ «Минск-32». Данный алгоритм может быть использован для некоторых задач анализа и синтеза фраз и предложений русского языка, в целях обнаружения и исправления ошибок в написании порядковых числительных и т. д.

ЛИТЕРАТУРА

1. Грамматика русского языка, т. I. М., Изд-во АН СССР, 1960. 719 с.
2. Гвоздев А. Н. Современный русский литературный язык. М., «Просвещение», 1967. 432 с.
3. Розенталь Д. Э. Русский язык. Пособие для поступающих в вузы. М., Изд-во Моск. ун-та, 1967. 303 с.
4. Орфографический словарь русского языка. М., Гос. изд-во иностр. и нац. словарей, 1959. 1259 с.
5. Добромыслов В. А., Розенталь Д. Э. Трудные вопросы грамматики и правописания. М., Учпедгиз, 1955. 288 с.
6. Шабанов-Кушнаренко Ю. П., Якименко Л. И. Об одной математической модели морфологической классификации множества имен существительных русского языка.— В сб.: Проблемы бионики. Вып. 6. Харьков, 1970, с. 104—107.

УДК 62. 506. 2

М. Ф. БОНДАРЕНКО, канд. техн. наук.,
Э. М. БУЗНИЦКАЯ, инж.

АЛГОРИТМ МОРФОЛОГИЧЕСКОГО АНАЛИЗА ИМЕН ПРИЛАГАТЕЛЬНЫХ РУССКОГО ЯЗЫКА

Рассмотрим моделирование речевого поведения человека при решении задач морфологического анализа имен прилагательных.

Задача 1. *Выделение классов имен прилагательных в соответствии с наличием противопоставлений по различным грамматическим категориям.*

Требуется произвести разбиение исходного множества M имен прилагательных русского языка на непересекающиеся подмножества (классы) M_1, M_2, M_3 по некоторому грамматическому показателю П, формирующемуся на основании сведений об окончании и характере основы. Каждое такое разбиение определяет между элементами этого множества отношение R типа эквивалентности, т. е. R является одновременно рефлексивным, симметричным и транзитивным.

Лингвистическими предпосылками организации формальных классов M_1, M_2, M_3 могут служить следующие соображения. В парадигме формального класса необходимо учитывать все противопоставления вариантов, свойственные членам этого класса. Поэтому парадигма может быть построена лишь на основании сопоставления парадигм отдельных лексем и выявления таким образом всех противопоставлений, встречающихся между элементами внутри данных частных парадигм.