УДК 62.506.2

А. Ф. ОСЫКА, Ю. П. ШАБАНОВ-КУШНАРЕНКО, д-р техн. наук

МОДЕЛЬ ОПРЕДЕЛЕНИЯ ИСХОДНОЙ ФОРМЫ ЧИСЛИТЕЛЬНЫХ РУССКОГО ЯЗЫКА

Целью данной работы является формальное описание, а также моделирование с помощью ЭВМ способности человека определять исходную форму русских числительных. В настоящем сообщении предлагается алгоритм, на вход которого поступают простые, сложные и составные количественные и порядковые числительные в любом падеже, записанные в соответствии с правилами орфографии русского языка. На выходе алгоритма получается исходная форма числительного, поступившего на вход. Под исходной формой мы понимаем числительное в именительном падеже. Если анализируемому числительному присущи категории рода и числа, то при определении исходной формы их значения не меняются.

Например, исходной формой числительного *ДВУХСОТ* является *ДВЕСТИ, ДВАДЦАТИ ОДНОЙ—ДВАДЦАТЬ ОДНА*. У числительного *ДВЕНАДЦАТОГО* имеется две исходных формы — *ДВЕНАДЦАТЫЙ* и *ДВЕНАДЦАТОЕ*. Понятие «исходной формы» в некоторых случаях синонимично понятию «форма наименования», которое встречается в американской лингвистической литературе. Однако в то время как форма наименования «...используется для того, чтобы говорить о всех членах парадигмы сразу», исходная форма является представителем подмножества словоформ парадигмы словоизменения с инвариантными значениями категорий рода и числа [1].

Таким образом, можно сказать, что на вход описываемого алгоритма подается последовательность $X=x_1, x_2, ..., x_i, ... x_n$ ($n=1, 2, ..., 23$), где x_i — падежная форма числительного, которое пишется одним словом и стоит на i-м месте во входной последовательности однословных числительных. В дальнейшем мы будем называть X входной словоформой. Если $n=1$, то на вход алгоритма поступило числительное, которое пишется одним словом. Если $n>1$, то на входе — составное числительное. Словоформы $x_1, x_2, ..., x_{n-1}$ всегда являются падежными формами количественных числительных. x_n может быть количественным либо порядковым.

На выходе алгоритма получается исходная форма числительного X, которую можно представить в виде последовательности $U=u_1, u_2, ..., u_i, ..., u_n$, где u_i — соответствующая падежная форма числительного x_i, которую мы будем называть выходной словоформой. u_i не всегда является числительным в именительном падеже, так как числительные *ТЫСЯЧА, МИЛЛИОН* и т. д. в составном числительном стоят в том числе и падеже, которого требует предыдущая словоформа [2].

Вся необходимая для работы предлагаемого алгоритма константная информация помещена в две таблицы, причем эта же информация используется для решения целого ряда задач по анализу и синтезу числительных русского языка. В графе № табл. 1 записаны номера основ количественных и порядковых числительных, помещенных в графе В. В графах Γ_1, Γ_2, Γ_3, Γ_4 записаны номера строк табл. 2, с окончаниями которых сочетается данная основа, если она является составной частью количественного числительного или стоит в препозиции в сложном порядковом числительном. В графах Γ_5, Γ_6, Γ_7, Γ_8 проставлены номера строк табл. 2, где записаны окончания, с которыми сочетается данная основа, если она входит в состав простого порядкового числительного или стоит в постпозиции в сложном порядковом числительном. В графе C_0 табл. 2 помещены номера строк (типов склонения) с окончаниями количественных и порядковых числительных. В строках 1—21 приводятся окончания количественных числительных, а в строках 22—34 — порядковых. В графах C_1, C_2, C_3, C_4, C_5, C_6 записаны окончания именительного, родительного и т. д. падежей соответствующих типов склонения. Незаполненные позиции на пересечении некоторых граф и строк табл. 2 обозначают так назы-

ваемое «нулевое окончание». Прочерки (----) указывают на то, что основы числительных, которые сочетаются с окончаниями данного типа склонения, не могут стоять в данном падеже.

Информация, помещенная в табл. 1 и 2, позволяет анализировать числительные, соответствующие числам натурального ряда в интервале от 1 до 10^{18}—1. Область определения данного алгоритма можно расширить за счет включения в табл. 1 редких названий чисел (секстильон и т. д.). Однако предлагаемый алгоритм при этом существенных изменений не потребует.

Описание алгоритма

0. Блок формальной проверки правильности входного сигнала [3, 4]. Перейти к п. 1.

1. Приступая к анализу очередной (начиная с первой) входной словоформы последовательности X, проверить, является ли она последней на входе. Если да, то перейти к п. 2, иначе — к п. 3.

2. Проверить, совпадает ли начало анализируемой словоформы с одной из основ (№ 27—33) табл. 1. Если да, то перейти к п. 4, иначе — к п. 3.

3. Найти в табл. 1 первую основу (№ 1—22), которая совпадает с началом анализируемой словоформы. Перейти к п. 4.

4. Вычеркнуть из анализируемой словоформы совпадающую часть и предшествующие ей буквы (если они имеются), запомнить номер основы, которая совпала с началом анализируемой словоформы. Перейти к п. 5.

5. Проверить, была ли вычеркнута в п. 4 из анализируемой словоформы одна из следующих основ: *ТЫСЯЧ, МИЛЛИОН, МИЛЛИАРД, ТРИЛЛИОН, КВАДРИЛЬОН*. Если да, то перейти к п. 8, иначе — к п. 6.

6. Проверить, является ли больше 26 номер вычеркнутой в п. 4 из анализируемой словоформы основы. Если да, то перейти к п. 35, иначе — к п. 7.

7. Проверить, содержит ли остаток P анализируемой словоформы после удаления в п. 4 выделенной основы больше трех букв. Если да, то перейти к п. 9, иначе — к п. 8.

8. Проверить, является ли анализируемая входная словоформа последней в последовательности X. Если да, то перейти к п. 35, если нет, то для выполнения действий в п. 21 запомнить номер графы Γ_1 табл. 1 и перейти к п. 19.

9. Проверить, совпадает одна из основ (№ 23—26) табл. 1 с остатком P анализируемой словоформы или его частью (совпадение должно произойти не обязательно начиная с первой буквы остатка P, но не позднее четвертой). Если да, то перейти к п. 11, иначе — к п. 10.

10. Найти в табл. 1 первую основу (№ 1—22), которая совпадает с остатком P анализируемой словоформы или его частью (совпадение должно произойти не обязательно начиная с первой буквы остатка P, но не позднее четвертой). Перейти к п. 4.

11. Вычеркнуть из остатка P анализируемой словоформы совпадающую часть и предшествующие ей буквы (если они имеются) и проверить, содержит ли новый остаток P более трех букв. Если да, то перейти к п. 10, иначе — к п. 12.

12. Проверить, является ли анализируемая входная словоформа последней во входной последовательности X. Если да, то перейти к п. 35, иначе — к п. 13.

13. Проверить, была ли в п. 9 выделена основа *НАДЦАТ*. Если да, то очередной выходной словоформой считать анализируемую входную словоформу без остатка P, полученного в п. 11, вместо которого приписать мягкий знак. Перейти к п. 42. Если нет, то перейти к п. 14.

14. Если в п. 3 была выделена основа *ВОСЬМ*, то в элемент памяти A записать *ВОСЕМ* и перейти к п. 16. В противном случае перейти к п. 15.

15. В элемент памяти A записать основу, выделенную в п. 3, а справа к ней приписать окончание, стоящее в табл. 2 на пересечении графы C_1 и строки, номер которой указан в табл. 1, в графе Γ_1 для основы, записанной в элемент памяти A. Перейти к п. 16.

Таблица 1

№	В	$Г_1$	$Г_2$	$Г_3$	$Г_4$	$Г_5$	$Г_6$	$Г_7$	$Г_8$
01	ДЕВЯНОСТ	08	—	—	—	23	24	25	34
02	СТ	08	—	—	—	—	—	—	—
03	ДЕСЯТ	06	—	—	—	23	24	25	34
04	СОРОК	09	—	—	—	—	—	—	—
05	ТЫСЯЧ	10	11	—	—	30	31	32	33
06	МИЛЛИОН	12	13	—	—	30	31	32	33
07	МИЛЛИАРД	12	13	—	—	30	31	32	33
08	ТРИЛЛИОН	12	13	—	—	30	31	32	33
09	КВАДРИЛЬОН	12	13	—	—	30	31	32	33
10	ДВАДЦАТ	06	—	—	—	23	24	25	34
11	ТРИДЦАТ	06	—	—	—	23	24	25	34
12	ДЕВЯТ	06	—	—	—	23	24	25	34
13	ВОСЕМ	21	—	—	—	—	—	—	—
14	ВОСЬМ	20	—	—	—	22	23	24	25
15	СЕМ	06	—	—	—	—	—	—	—
16	ШЕСТ	06	—	—	—	22	23	24	25
17	ПЯТ	06	—	—	—	23	24	25	34
18	ЧЕТЫР	17	—	—	—	—	—	—	—
19	ТР	16	—	—	—	—	—	—	—
20	ДВ	15	14	—	—	—	—	—	—
21	ОДИН	01	—	—	—	—	—	—	—
22	ОДН	02	03	05	04	—	—	—	—
23	НАДЦАТ	06	—	—	—	23	24	25	34
24	ДЕСЯТ	07	—	—	—	23	24	25	34
25	СОТ	18	—	—	—	23	24	25	34
26	СТ	19	—	—	—	—	—	—	—
27	СОРОКОВ	—	—	—	—	22	23	24	25
28	ТРЕТ	—	—	—	—	26	27	28	29
29	ЧЕТВЕРТ	—	—	—	—	23	24	25	34
30	СЕДЬМ	—	—	—	—	22	23	24	25
31	ВТОР	—	—	—	—	22	23	24	25
32	ПЕРВ	—	—	—	—	23	24	25	34
33	СОТ	—	—	—	—	23	24	25	34

16. Проверить, была ли выделена в п. 9 из состава анализируемой входной словоформы основа *ДЕСЯТ*. Если да, то считать очередной выходной словоформой содержимое элемента памяти *А*, к которому справа приписать *ДЕСЯТ* и перейти к п. 42. В противном случае перейти к п. 17.

17. Проверить, была ли выделена в п. 3 из состава анализируемой словоформы основа *ДВ*. Если да, то считать очередной выходной словоформой содержимое элемента памяти *А*, к которому справа приписать *СТИ* и перейти к п. 42. В противном случае перейти к п. 18.

18. Считать очередной выходной словоформой содержимое элемента памяти *А*, к которому справа приписать *СТА*, если в п. 3 из состава анализируемой словоформы была выделена основа *ТР* или *ЧЕТЫР*. В противном случае приписать *СОТ*. Перейти к п. 42.

19. Проверить, была ли выделена в п. 3 основа *ТЫСЯЧ*. Если да, то перейти к п. 32, иначе — к п. 20.

20. Проверить, была ли выделена в п. 3 одна из следующих основ: *МИЛЛИОН, МИЛЛИАРД, ТРИЛЛИОН, КВАДРИЛЬОН*. Если да, то перейти к п. 24, иначе — к п. 21.

21. Проверить, равно ли прочерку (-----) окончание в табл. 2, стоящее на пересечении графы $С_1$ и строки, номер которой записан в табл. 1 для выделенной в п. 3 основы, в ранее запомненной графе. Если да, то перейти к п. 22, иначе — к п. 23.

22. В элемент памяти *А* записать основу, у которой порядковый номер на единицу меньше, чем у выделенной в п. 3 основы. В качестве очередной выходной словоформы взять содержимое элемента памяти *А*, к которому справа приписать окончание, стоящее в табл. 2 на пересечении графы C_1 и строки, номер которой помещен в табл. 1 в графе $Г_1$ для записанной в элемент памяти *А* основы. Перейти к п. 42.

23. В качестве очередной выходной словоформы взять основу, выделенную в п. 3, к которой справа приписать окончание, полученное в п. 21. Перейти к п. 42.

24. Проверить, является ли анализируемая словоформа первой во входной последовательности *X*. Если да, то перейти к п. 25, иначе — к п. 26.

25. Запомнить номер графы $Г_1$ в табл. 1 и перейти к п. 21.

26. Проверить, была ли выделена последней из состава предыдущей входной словоформы основа *ОДИН* или основа *ОДН*. Если да, то вместо ранее полученной предыдущей выходной словоформы записать числительное *ОДИН* и перейти к п. 25. Если нет, то перейти к п. 27.

27. Проверить, была ли выделена последней из состава предыдущей входной словоформы основа *ДВ*. Если да, то вместо ранее полученной предыдущей выходной словоформы записать числительное *ДВА* и перейти к п. 28. Если нет, то перейти к п. 29.

28. В качестве очередной выходной словоформы взять выделенную в п. 3 основу и справа к ней приписать окончание, стоящее в табл. 2 на пересечении графы C_2 и строки, номер которой записан в графе $Г_1$ таблицы 1 для выделенной в п. 3 основы. Перейти к п. 42.

29. Проверить, была ли выделена последней из состава предыдущей входной словоформы основа *ТР* или основа *ЧЕТЫР*. Если да, то перейти к п. 2[illegible], иначе — к п. 30.

30. Проверить, была ли выделена из состава предыдущей входной словоформы одна из следующих основ: *ТЫСЯЧ, МИЛЛИОН, МИЛЛИАРД, ТРИЛЛИОН, КВАДРИЛЬОН*. Если да, то перейти к п. 25. В противном случае перейти к п. 31.

31. В качестве очередной выходной словоформы взять выделенную в п. 3 основу и справа к ней приписать окончание, стоящее в табл. 2 на пересечении графы C_2 и строки, номер которой записан в графе $Г_2$ табл. 1 для выделенной в п. 3 основы. Перейти к п. 42.

32. Проверить, является ли анализируемая словоформа первой во входной последовательности *X*. Если да, то перейти к п. 25, иначе — к п. 33.

33. Проверить, была ли выделена последней из состава предыдущей входной словоформы основа *ОДН*. Если да, то вместо ранее полученной предыдущей выходной словоформы записать числительное *ОДНА* и перейти к п. 2[illegible]. Если нет, то перейти к п. 34.

34. Проверить, была ли выделена последней из состава предыдущей словоформы основа *ДВ*. Если да, то перейти к п. 28, иначе — к п. 29.

35. Проверить, совпадает ли остаток *Р* анализируемой словоформы с окончаниями, записанными в тех строках табл. 2, номера которых указаны в табл. 1 в графах $Г_1$, $Г_2$, $Г_3$, $Г_4$, $Г_5$, $Г_6$, $Г_7$, $Г_8$ для основы, выделенной последней из состава анализируемой входной словоформы. Если совпадение произошло, то запомнить номера строк, которые содержат окончания, совпавшие с остатком *Р*. Перейти к п. 37. В случае отсутствия совпадения перейти к п. 36.

36. Проверить, является ли больше 22 номер последней основы, выделенной из состава анализируемой словоформы. Если да, то перейти к п. 13, иначе — к п. 19.

37. Проверить, является ли больше 21 номер очередной (начиная с первой установленной в п. 35) строки табл. 2, содержащей окончания, совпавшие с остатком *Р*. Если да, то перейти к п. 38. Если нет, то запомнить номер графы в табл. 1, в которой записан номер данной строки табл. 2, и перейти к п. 3[illegible].

38. В качестве варианта последней выходной словоформы u_n взять анализируемую входную словоформу без остатка *Р*, к которой справа приписать окончание, стоящее в табл. 2 на пересечении графы C_1 и строки, которая проверялась в п. 37. Перейти к п. 39.

Таблица 2

C_0	C_1	C_2	C_3	C_4	C_5	C_6
01		– – –	– – –		– – –	– – –
02	– – –	ого	ому	– – –	им	ом
03	а	ой	ой	у	ой	ой
04	о	ого	ому	о	им	ом
05	и	их	им	и	ими	их
06	ь	и	и	ь	ью	и
07		и	и		ью	и
08	о	а	а	о	а	а
09		а	а		а	а
10	а	и	е	у	ей	е
11	и		ам	и	ами	ах
12		а	у		ом	е
13	ы	ов	ам	ы	ами	ах
14	а	ух	ум	а	умя	ух
15	е	ух	ум	е	умя	ух
16	и	ех	ем	и	емя	ех
17	е	ех	ем	е	ьмя	ех
18			– – –		– – –	– – –
19	и	– – –	ам	и	ами	ах
20	– – –	и	и	– – –	ью	и
21	ь	– – –	– – –	ь	ью	– – –
22	ой	ого	ому	ой	ым	ом
23	ая	ой	ой	ую	ой	ой
24	ое	ого	ому	ое	ым	ом
25	ые	ых	ым	ые	ыми	ых
26	ий	ьего	ьему	ий	ьим	ьем
27	ья	ьей	ьей	ью	ьей	ьей
28	ье	ьего	ьему	ье	ьим	ьем
29	ьи	ьих	ьим	ьи	ьими	ьих
30	ный	ного	ному	ный	ным	ном
31	ная	ной	ной	ную	ной	ной
32	ное	ного	ному	ное	ным	ном
33	ные	ных	ным	ные	ными	ных
34	ый	ого	ому	ый	ым	ом

39. Выдать на печать последовательность сформированных выходных словоформ — $U=u_1, u_2,..., u_i, ..., u_n$, которая является исходной формой числительного на входе. Перейти к п. 40.

40. Проверить, была ли сформирована последняя выходная словоформа в одном из следующих пунктов данного алгоритма: п. 16, п. 17, п. 18. Если да, то перейти к п. 43, иначе — к п. 41.

41. Проверить, был ли сформирован вариант выходной словоформы для строки табл. 2, которая была заполнена последней в п. 35. Если да, то перейти к п. 43, иначе — к п. 37.

42. Проверить, была ли сформирована выходная словоформа, соответствующая последней входной словоформе. Если да, то перейти к п. 39, иначе — к п. 1.

43. Конец работы алгоритма.

Данный алгоритм был реализован на ЭВМ «Минск-32». Во всех проведенных экспериментах были получены удовлетворительные результаты, которые могут быть использованы для решения задач анализа и синтеза текстов на русском языке: для автоматической корректировки текстов, автоматического реферирования и т. п.

СПИСОК ЛИТЕРАТУРЫ

1. Хэмп Э. Словарь американской лингвистической терминологии. М., «Прогресс», 1964. 264 с.