УДК 519.766.2

М. Ф. БОНДАРЕНКО, канд. техн. наук, *В. М. БОНДАРЕВ*

О МАТЕМАТИЧЕСКОМ ОПИСАНИИ СЛОВОИЗМЕНЕНИЯ СУЩЕСТВИТЕЛЬНЫХ. *СООБЩЕНИЕ 1*

Описание любого фрагмента морфологической системы языка можно дать в виде предиката $L(X, Y, Z)$, где предметной областью X является множество слов (словарных форм), предметной областью Y — множество словоформ, а Z — множество значений грамматических категорий, характеризующих словоформу Y [1]. Предикат $L(X, Y, Z)$ принимает значение «истина», если значения переменных не являются взаимоисключающими с точки зрения нормы языка. В противном случае значение предиката — «ложь». Обозначим «истину» и «ложь» 1 и 0 соответственно, а сам предикат назовем морфологической функцией [1].

Если для морфологической функции существует аналитическое выражение, то разнообразные задачи морфологической обработки, в том числе синтез, анализ, нормализацию, можно представить уравнениями типа $L(X, Y, Z) = 1$, где, в зависимости от конкретной задачи, значения тех или иных переменных известны. Решение всякой задачи сведется при этом к поиску корней соответствующего уравнения.

Цель настоящей работы — предложить возможный способ описания предиката $L(X, Y, Z)$ и проиллюстрировать этот способ формализацией небольшого фрагмента морфологии русского языка.

В традиционном описании морфологии важную роль играют такие понятия, как основа слова, его окончание, тип склонения, чередование в основе и т. п. Всю совокупность таких понятий условно назовем морфологической характеристикой и обозначим Г. Формально Г — это набор переменных, каждая из которых принимает конечное число значений.

Из практики языка следует, что всякому совместимому набору значений переменных X, Y, Z можно поставить в соответствие хотя бы один совместимый с ним набор значений переменных Г. Поэтому можно задать предикат $L'(X, Y, Z, \Gamma)$, принимающий значение 1, когда все переменные X, Y, Z и Г совместимы, и 0 — в противном случае. Поскольку решение уравнений есть поиск таких значений неизвестных величин, которые были бы совместимы со значениями величин известных, ясно, что, приняв заданными некоторые переменные из X, Y, Z и решив уравнение $L'(X, Y, Z, \Gamma) = 1$, отыщем значения неизвестных переменных из набора X, Y, Z, так как если бы решали уравнение $L(X, Y, Z) = 1$. Другими словами, имея аналитическое выражение для морфологической функции L', можно решить все те задачи, которые решили бы, имея выражение для L.

Целесообразно формализовать предикат L', а не L, так как в этом случае можно полнее использовать существующие описания морфологии, например грамматику русского языка.

Заранее оговорим, что выражение для L' будем искать в виде конъюнкции предикатов L_1, L_2, ..., L_k от тех же переменных: X, Z, Г. Предпосылку к такому решению видим в том, что морфология обычно описывается набором параграфов или правил, совместное выполнение которых обеспечивает правильность грамматической обработки.

Процесс формализации морфологии представляется нам в виде накопления формул, выражающих грамматические правила, записанные в форме предикатов L_1, L_2, ..., L_k. Критерием качества математического описания может быть статистический эксперимент, показывающий, насколько удовлетворяют норме языка решения морфологических задач, полученные с помощью модели вида $L_1 \wedge L_2 \wedge \ldots \wedge L_k$.

Хотя формально всякий предикат L_i ($i = 1, 2, \ldots, k$) зависит от всех своих переменных, его фактическими аргументами является лишь часть их, для каждого L_i — своя. Остальные переменные можно считать связанными кванторами общности, которые в записи формул опускаем.

Чтобы придать содержательность нашим построениям, рассмотрим конкретный пример грамматического описания, а именно, описание склонения существительных среднего рода, оканчивающихся на *-е*, таких как *море, поле, солнце*. Пример этот достаточно прост и может быть разобран в рамках настоящей статьи, в то же время он отражает многие особенности слово-

изменения существительных. Заранее условимся, что слова и словоформы будут записаны без знака ударения. Исключим из рассмотрения существительные, склоняющиеся по типу прилагательных, например *животное*, а также те слова с дефисом, у которых склоняются обе части.

Поскольку нас интересует письменная форма языка, X и Y удобно представлять в виде наборов переменных $X = <x_1, x_2, \ldots, x_n>$, $Y = <y_1, y_2, \ldots, y_n>$. Каждая переменная из этих наборов соответствует отдельной позиции в записи слова или словоформы, значением переменной является буква или другой символ, стоящий в этой позиции. Число n должно быть выбрано с таким расчетом, чтобы вместить слово или словоформу максимальной длины.

Будем считать, что переменная Z также представляет собой набор $<z_1, z_2>$, где z_1 — грамматическая категория числа, а z_2 — категория падежа. Понятно, что при описании другого фрагмента морфологии этот набор может быть иным, в частности более полным. В качестве морфологической характеристики выберем набор переменных $\Gamma = <\alpha, \beta, \gamma, \eta, \omega>$. Здесь α — основа слова; β — чередование в основе словоформы; γ — тип склонения; η — основа словоформы; ω — окончание словоформы.

Так как у нас имеется фиксированное число переменных $x_1, x_2, \ldots, x_n$, а длина слов различна, создается противоречие, разрешить которое можно, дополняя каждое слово особыми символами до стандартной длины. Будем использовать для этой цели символ пробела ␣, а само слово размещать в крайних левых позициях. Так, если $X = поле$, $n = 6$, то $x_1 = $ п, $x_2 = $ о, $x_3 = $ л, $x_4 = $ е, $x_5 = $ ␣, $x_6 = $ ␣. То же касается и переменных $y_i (i = 1, 2, \ldots, n)$.

Таким образом, область определения каждой из упомянутых переменных состоит из букв русского алфавита, дефиса и знака пробела — всего 35 символов. Это дает нам основание записать предикаты

$$x_i^{а} \vee x_i^{б} \vee \ \ldots \ \vee x_i^{я} \vee x_i^{-} \vee x_i^{␣} \quad (i = 1, 2, \ldots, n); \qquad (1)$$

$$y_i^{а} \vee y_i^{б} \vee \ \ldots \ \vee y_i^{я} \vee y_i^{-} \vee y_i^{␣} \quad (i = 1, 2, \ldots, n), \qquad (2)$$

которые с помощью символики, принятой в работах [1; 2], выражают тот факт, что значением переменных x_i и y_i могут быть лишь определенные символы.

Области определения переменных z_1 и z_2 состоят соответственно из двух значений числа {е, м} и шести значений падежа {и, р, д, в, т, п};

$$z_1^{е} \vee z_1^{м}; \qquad (3)$$

$$z_2^{и} \vee z_2^{р} \vee z_2^{д} \vee z_2^{в} \vee z_2^{т} \vee z_2^{п}. \qquad (4)$$

В нашем конкретном примере под основой слова будем понимать часть слова, оставшуюся после удаления последней буквы *е*, если слово относится к разряду склоняемых, и слово целиком, если оно не склоняется. Поскольку α — внутренняя переменная, т. е. не может быть ни входной ни выходной при решении задач синтеза, анализа, нормализации, в целях упрощения описания обозначим всякую основу просто числом, так называемым номером основы, при этом разные основы будут иметь различные номера. Областью определения переменной α будем считать множество чисел $\{1, 2, \ldots, t\}$, где t — общее число основ рассматриваемого класса слов,

$$\alpha^1 \vee \alpha^2 \vee \ldots \vee \alpha^t. \qquad (5)$$

То же самое можно сказать о переменной ω — окончании словоформы. В отличие от α число ее значений вполне обозримо и составляет 18 элементов:

$$\omega^1 \vee \omega^2 \vee \ldots \vee \omega^{18}. \qquad (6)$$

Раскрывая смысл переменной β, заметим, что в рамках рассматриваемого примера имеют место лишь двуступенчатые чередования [3], например: *поленце — поленец*, *ущелье — ущелий*. Отсюда следует, что область определения переменной β можно ограничить двумя символами $\{1, 2\}$, где 1 будет означать первую ступень чередования, а 2 — вторую,

$$\beta^1 \vee \beta^2. \qquad (7)$$

Переменная γ представляет тип склонения существительного, который будем понимать так же, как в работе [4].

Оттуда же будем черпать всю необходимую лингвистическую информацию. Хотя в [4] отмечается 9 возможных типов склонения существительных, для нашего примера актуальны лишь 6 из них: 0 — несклоняемые существительные; 2 — существительные стандартного мягкого склонения; 4 — существительные с основой на шипящую; 5 — существительные с основой на *ц*; 6 — с основой на гласную (кроме *и*), *ь* или *й*; 7 — с основой на букву *и*. Множество 0, 2, 4, 5, 6, 7 будем считать областью изменения переменной γ:

$$\gamma^0 \vee \gamma^2 \vee \gamma^4 \vee \gamma^5 \vee \gamma^6 \vee \gamma^7. \qquad (8)$$

Основой словоформы η будем считать часть словоформы без окончания, дополненную пробелами до стандартной длины. В отличие от основы слова η представим не числом, а набором переменных $(\eta_1, \eta_2, \ldots, \eta_n)$ подобно тому, как представлены слово X и форма Y. Область изменения переменных $y_i (i = 1, 2, \ldots, n)$ такая же, как переменных x_i и y_i,

$$\eta_i^{\text{а}} \vee \eta_i^{\text{б}} \vee \ldots \vee \eta_i^{\text{я}} \vee \eta_i^{\bar{}} \vee \eta_i^{\sqcup} \quad (i = 1, 2, \ldots, n). \qquad (9)$$

Центральное место в описании склонения занимает предикат, связывающий грамматические категории числа и падежа с типом склонения и падежными окончаниями. Обозначим его $L_1(\gamma, z_1, z_2, \omega)$ или просто L_1. В грамматике принято отдельно описывать словоизменение существительных каждого типа склонения. Можно передать это так: если слово имеет нулевой тип склонения, то связь падежей, чисел и окончаний описывается предикатом $L_1(0, z_1, z_2, \omega)$, если второй тип склонения — предикатом $L_1(2, z_1, z_2, \omega)$ и т. д.,— перебирая все необходимые типы склонения. Переведем эту фразу на язык формального исчисления:

$$L_1 = \bigwedge_{i=0,\,2,\,4,\,5,\,6,\,7} \gamma^i \supset L_1(i, z_1, z_2, \omega), \tag{10}$$

где $\wedge$ означает логическое произведение по всем значениям переменной i, перечисленным под ним.

Известно, что описание каждого типа склонения складывается из описаний склонения в единственном и во множественном числе, т. е. в i-м типе склонения в единственном числе окончания и падежи связаны предикатом $L_1(i, e, z_2, \omega)$, а во множественном числе — предикатом $L_1(i, м, z_2, \omega)$:

$$L_1(i, z_1, z_2, \omega) = (z_1^{\text{е}} \subset L_1(i, \text{е}, z_2, \omega))(z_1^{\text{м}} \supset \supset L_1(i, \text{м}, z_2, \omega)). \tag{11}$$

Всего имеется 6 выражений вида (11), для каждого типа склонения свое. В записи формулы (11) опущен знак операции конъюнкции, который должен стоять между двумя сомножителями. Будем опускать его везде, где это не вызовет недоразумения.

Рассмотрим склонение единственного числа существительных стандартного (второго) типа. Известно, что если такие существительные стоят в именительном, винительном и предложном падеже, то окончание их *-е*, в родительном *-я*, в дательном *-ю*, в творительном *-ем*:

$$L_1(2, \text{е}, z_2, \omega) = (z_2^{\text{и}} \vee z_2^{\text{в}} \vee z_2^{\text{п}} \supset \omega^{\text{е}})(z_2^{\text{р}} \supset \supset \omega^{\text{я}})(z_2^{\text{д}} \supset \omega^{\text{ю}})\, z_2^{\text{т}} \supset \omega^{\text{ем}}). \tag{12}$$

Хотя мы условились всякое окончание обозначать числом, здесь и ниже будем пользоваться его буквенной записью, чтобы не лишать изложение наглядности. Эту запись можно толковать как код числа в 33-ичной системе счисления.

Склонение слов стандартного типа во множественном числе определяется следующим правилом. Если падеж именительный, то окончание слова *-я*; если родительный, то в случае падения ударения на основу слово оканчивается на *-ь*, при ударении на окончании — на *-ей*; если падеж дательный, то окончание *-ям*; если падеж винительный, то окончание неодушевленных существительных такое же, как окончание существительных в имени-

тельном падеже, а одушевленных — такое же, как в родительном; если падеж творительный, то окончание *-ями*; если предложный, *-ях*. Рассмотрим отдельно ту часть правила, которая гласит: если падеж родительный, то в случае падения ударения на основу слово оканчивается на *-ь*, при ударении на окончании — на *-ей*.

У нас нет средств адекватно формализовать ее смысл, так как нет переменной, которая каким-то образом учитывала бы место ударения. Поэтому заменим эту часть правила ее неточным аналогом: если падеж родительный, то окончание словоформы *-ь* или *-ей*, что, конечно, огрубляет описание в целом. То же касается фрагмента правила, описывающего окончания винительного падежа. Так как одушевленность или неодушевленность также не учитываются, заменим этот фрагмент следующим: если падеж винительный, то окончание совпадает с окончанием именительного или родительного падежа. После этих замен правило в целом выразится следующим предикатом:

$$L_1(2,\ \text{м},\ z_2,\ \omega) = (z_2^{\text{и}} \supset \omega^{\text{я}})\,(z_2^{\text{р}} \supset \omega^{\text{ь}} \vee \omega^{\text{ей}})\,(z_2^{\text{д}} \supset \omega^{\text{ям}}) \wedge$$
$$\wedge\,(z_2^{\text{в}} \supset \omega^{\text{я}} \vee \omega^{\text{ь}} \vee \omega^{\text{ей}})\,(z_2^{\text{т}} \supset \omega^{\text{ями}})\,(z_2^{\text{п}} \supset \omega^{\text{ях}}). \qquad (13)$$

Рассмотрим нулевой тип склонения, который значительно отличается от прочих типов. Известно, что слова нулевого типа имеют пустое окончание во всех падежах и числах, т. е.

$$L_1(0,\ z_1,\ z_2,\ \omega) = \omega^{\sqcup}, \qquad (14)$$

где $\sqcup$ в данном случае символизирует пустое окончание. Для остальных типов склонения приведем соответствующие формулы:

$$L_1(4,\ \text{е},\ z_2,\ \omega) = (z_2^{\text{и}} \vee z_2^{\text{в}} \vee z_2^{\text{п}} \supset \omega^{\text{е}})\,(z_2^{\text{р}} \supset \omega^{\text{а}}) \wedge$$
$$\wedge\,(z_2^{\text{д}} \supset \omega^{\text{у}})\,(z_2^{\text{т}} \supset \omega^{\text{ем}}); \qquad (15)$$

$$L_1(4,\ \text{м},\ z_2,\ \omega) = (z_2^{\text{и}} \supset \omega^{\text{а}})\,(z_2^{\text{р}} \supset \omega^{\sqcup})\,(z_2^{\text{д}} \supset \omega^{\text{ам}}) \wedge$$
$$\wedge\,(z_2^{\text{в}} \supset \omega^{\text{а}} \vee \omega^{\sqcup})\,(z_2^{\text{т}} \supset \omega^{\text{ами}})\,(z_2^{\text{п}} \supset \omega^{\text{ах}}); \qquad (16)$$

$$L_1(5,\ \text{е},\ z_2,\ \omega) = L_1(4,\ \text{е},\ z_2,\ \omega); \qquad (17)$$

$$L_1(5,\ \text{м},\ z_2,\ \omega) = L_1(4,\ \text{м},\ z_2,\ \omega); \qquad (18)$$

$$L_1(6,\ \text{е},\ z_2,\ \omega) = L_1(2,\ \text{е},\ z_2,\ \omega); \qquad (19)$$

$$L_1(6,\ \text{м},\ z_2,\ \omega) = (z_2^{\text{и}} \supset \omega^{\text{я}})\,(z_2^{\text{р}} \supset \omega^{\text{й}})\,(z_2^{\text{д}} \supset \omega^{\text{ям}}) \wedge$$
$$\wedge\,(z_2^{\text{в}} \supset \omega^{\text{я}} \vee \omega^{\text{й}})\,(z_2^{\text{т}} \supset \omega^{\text{ями}})\,(z_2^{\text{п}} \supset \omega^{\text{ях}}); \qquad (20)$$

$$L_1(7,\ \text{е},\ z_2,\ \omega) = (z_2^{\text{и}} \vee z_2^{\text{в}} \supset \omega^{\text{е}})\,(z_2^{\text{р}} \supset \omega^{\text{я}})\,(z_2^{\text{д}} \supset \omega^{\text{ю}}) \wedge$$
$$\wedge\,(z_2^{\text{т}} \supset \omega^{\text{ем}})\,(z_2^{\text{п}} \supset \omega^{\text{и}}); \qquad (21)$$

$$L_1(7,\ \text{м},\ z_2,\ \omega) = (z_2^{\text{и}} \supset \omega^{\text{я}})\,(z_2^{\text{р}} \supset \omega^{\text{й}} \vee \omega^{\text{ев}})\,(z_2^{\text{д}} \supset \omega^{\text{ям}}) \wedge$$
$$\wedge\,(z_2^{\text{в}} \supset \omega^{\text{я}} \vee \omega^{\text{й}})\,(z_2^{\text{т}} \supset \omega^{\text{ями}})\,(z_2^{\text{п}} \supset \omega^{\text{ях}}). \qquad (22)$$

Важную роль в математической модели играют предикаты, связывающие слово X с морфологическими характеристиками α и γ. В нашем примере каждой из основ соответствует слово, и притом только одно:

$$L_2(X, \alpha) = \bigwedge_{i=1}^{t} (\alpha \sim x_1^{a_{i1}} x_2^{a_{i2}} \ldots x_n^{a_{in}}). \quad (23)$$

Здесь буквы a_{i1}, a_{i2}, ..., a_{in} составляют слово с основой i.

Распределение слов по типам склонения можно описать следующим образом. Если предпоследняя буква слова *л* или *р*, то слово 2-го типа склонения; *ж*, *ч* или *щ*, то 4-го; если *ц*, то 5-го; если *ь*, то 6-го; если *и*, то 7-го. Для формализации этого правила воспользуемся промежуточной переменной L_{31}, которой обозначим предпоследнюю букву слова:

$$L_{31} = (\mu^{л} \vee \mu^{р} \supset \gamma^2)(\mu^{ж} \vee \mu^{ч} \vee \mu^{щ} \supset \gamma^4)(\mu^{ц} \supset \gamma^5)(\mu^{ь} \supset \gamma^6)(\mu^{и} \supset \gamma^7). \quad (24)$$

К этому следует добавить, что на предпоследнем месте у склоняемых существительных среднего рода, кончающихся на *-е*, возможна лишь одна из букв: *л*, *р*, *ж*, *ч*, *щ*, *ц*, *ь*, *и*:

$$L_{32} = \mu^{л} \vee \mu^{р} \vee \mu^{ж} \vee \mu^{ч} \vee \mu^{щ} \vee \mu^{ц} \vee \mu^{ь} \vee \mu^{и}. \quad (25)$$

Само же понятие «предпоследняя буква» задается предикатом

$$L_{33} = (\forall i \in \{1, 2, \ldots, n-2\})[x_{i+1}^{е} x_{i+2}^{\square} \supset (\mu = x_i)]. \quad (26)$$

Формула $\mu = x_i$ является сокращенной записью выражения

$$\mu^{а} x_i^{а} \vee \mu^{б} x_i^{б} \vee \ldots \vee \mu^{я} x_i^{я} \vee \mu^{\bar{}} x_i^{\bar{}} \vee \mu^{\square} x_i^{\square}.$$

Правило, которое выражается конъюнкцией предикатов L_{31}, L_{32}, L_{33}, справедливо только для склоняемых существительных. Тип склоняемых несклоняемых существительных — нулевой, он не зависит от того, какая буква стоит на предпоследнем месте:

$$L_{34} = \gamma^0. \quad (27)$$

Все несклоняемые существительные полностью определяются перечнем своих основ — M_0. При этом общее правило определения типа склонения выглядит так:

$$L_3(X, \alpha, \gamma) = (\overline{\alpha \in M_0} \supset L_{31} L_{32} L_{33})(\alpha \in M_0 \supset L_{34}). \quad (28)$$

Выражение $\alpha \in M_0$ представляет собой сокращенную запись предиката $\alpha^{m_1} \vee \alpha^{m_2} \vee \ldots \vee \alpha^{m_k}$, где $\{m_1, m_2, \ldots, m_k\} = M_0$.

Список литературы: 1. *Шабанов-Кушнаренко Ю. П.* Применение метода нульоргана в лингвистике.— Проблемы бионики. Харьков, 1978, вып. 21, с. 3—15. 2. *Шабанов-Кушнаренко Ю. П.* О теории интеллекта.— Проблемы бионики. Харьков, 1979, вып. 22. с. 3—13. 3. *Зализняк А. А.* Русское именное словоизменение. М., Наука, 1967. 370 с. 4. *Зализняк А. А.* Грамматический словарь русского языка. М., Русский язык, 1977. 879 с.