УДК 519.766.2

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

О МАТЕМАТИЧЕСКОМ ОПИСАНИИ СЛОВОИЗМЕНЕНИЯ СУЩЕСТВИТЕЛЬНЫХ. *СООБЩЕНИЕ 2*

Необходимой частью математического описания словоизменения существительных являются предикаты, связывающие словоформу с морфологическими характеристиками. Представляется полезным уметь выделить буквы основы словоформы и указать ее окончание.

Рассмотрим следующее высказывание. Если окончание словоформы *-я*, то существует такой номер переменной k, что $y_i = \eta_i (i < k)$, $y_k = я$, $\eta_k = \sqcup$, а все переменные y_i и $\eta_i (i > k)$ равны пробелу,

$$L_{4,я} = \omega^{я} \supset (\exists k \in \{1, 2, \ldots, n-1\}) \left[\bigwedge_{i=1}^{k-1} (y_i = \eta_i) \overline{\eta_{k-1}^{\sqcup}} \wedge \right.$$

$$\left. \wedge y_k^{я} \eta_k^{\sqcup} \bigwedge_{i=k+1}^{n} y_i^{\sqcup} \eta_i^{\sqcup} \right]. \qquad (29)$$

Запись $(\exists k \in \{1, 2, \ldots, n\}) P(k)$ означает $P(1) \vee P(2) \vee \ldots \vee P(n)$.

Трехбуквенному окончанию *-ями* соответствует предикат:

$$L_{4,\text{ями}} = \omega^{\text{ями}} \supset (\exists k \in \{1, 2, \ldots, n-3\}) \left[\bigwedge_{i=1}^{k-1} (y_i = \eta_i) \overline{\eta_{k-1}^{\sqcup}} \wedge \right.$$

$$\left. \wedge y_k^{я} y_{k+1}^{м} y_{k+2}^{и} \eta_k^{\sqcup} \eta_{k+1}^{\sqcup} \eta_{k+2}^{\sqcup} \bigwedge_{i=k+3}^{n} y_i^{\sqcup} \eta_i^{\sqcup} \right]. \qquad (30)$$

Обобщим выражения типа (29) и (30):

$$L_{4,i} = \omega^{i} \supset (\exists k \in \{1, 2, \ldots, n-3\}) \left[\bigwedge_{i=1}^{k-1} (y_i = \eta_i) \overline{\eta_{k-1}^{\sqcup}} \wedge \right.$$

$$\left. \wedge y_k^{a_1} y_{k+1}^{a_2} y_{k+2}^{a_3} \eta_k^{\sqcup} \eta_{k+1}^{\sqcup} \eta_{k+2}^{\sqcup} \bigwedge_{i=k+3}^{n} y_i^{\sqcup} \eta_i^{\sqcup} \right]. \qquad (31)$$

Каждому значению переменной i соответствует один столбец табл. 1.

Таблица 1

ω	е	я	а	ю	у	и	й	ь	ем	ей	ев	ям	ам	ях	ах	ями	ами	$\sqcup$
a_1	е	я	а	ю	у	и	й	ь	е	е	е	я	а	я	а	я	а	$\sqcup$
a_2	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	м	й	в	м	м	х	х	м	м	$\sqcup\sqcup$
a_3	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	$\sqcup$	и	и	$\sqcup$

Если окончание словоформы пустое, выражение (31) можно упростить:

$$L_{4,\sqcup} = \omega^{\sqcup} \supset \bigwedge_{i=1}^{n} (y_i = \eta_i). \tag{32}$$

Общий предикат, учитывающий связь переменных y, ω и η, будет логическим произведением всевозможных предикатов $L_{4,i}$:

$$L_4(y, \omega, \eta) = \bigwedge_{i=1}^{18} L_{4,i}. \tag{33}$$

Изучим предикаты, описывающие чередование в основах словоформ. Рассмотрим связь между словом X, основой словоформы η, типом склонения γ и ступенью чередования β. Если в словоформе имеет место первая ступень чередования, то все буквы слова X и основы η, совпадают, начиная с 1-й и кончая некоторой k-й буквой, причем эта буква — последняя в основе, не являющаяся пробелом. Если тип склонения нулевой, то $(k+1)$-я буква слова — $\sqcup$, если не нулевой, то $(k+1)$-я буква слова — e. Все остальные буквы основы и слова — пробелы:

$$L_{5,1} = \beta^1 \supset (\exists k \in \{1, 2, \ldots, n-2\}) \left[\bigwedge_{i=1}^{k} (\eta_i = x_i) \overline{\eta_k^{\sqcup}} \wedge \right.$$
$$\left. \wedge \eta_{k+1}^{\sqcup} (\gamma^0 x_{k+1}^{\sqcup} \vee \overline{\gamma^0} x_{k+1}^{e}) \bigwedge_{i=k+2}^{n} x_i^{\sqcup} \eta_i^{\sqcup} \right]. \tag{34}$$

Таблица 2

x_k	x_{k+1}	x_{k+2}	η_k	η_{k+1}	η_{k+2}
ь	е	$\sqcup$	и	$\sqcup$	$\sqcup$
в	ц	е	в	е	ц
д	ц	е	д	е	ц
н	ц	е	н	е	ц
р	ц	е	р	е	ц
т	ц	е	т	е	ц
ь	ц	е	е	ц	$\sqcup$

Если ступень чередования вторая, то соответствие между буквами слова и основы нагляднее всего выразится табл. 2.

Все буквы слова и основы словоформы с 1-й по $(k-1)$-ю попарно одинаковы, а все символы, следующие за $(k+2)$-м, являются пробелами. После вынесения за скобки общих элементов предикат, описывающий таблицу и замечания к ней, будет следующим:

$$L_{5,2} = \beta^2 \supset (\exists k \in \{1, 2, \ldots, n-3\}) \left[\left(\bigwedge_{i=1}^{k-1} x_i = \eta_i \bigwedge_{i=k+3}^{n} x_i^{\sqcup} \eta_i^{\sqcup} \right) \wedge \right.$$
$$\wedge (x_k^{\text{ь}} x_{k+1}^{\text{е}} x_{k+2}^{\sqcup} \eta_k^{\text{и}} \eta_{k+1}^{\sqcup} \eta_{k+2}^{\sqcup} \vee x_{k+1}^{\text{ц}} x_{k+2}^{\text{е}} (\eta_{k+1}^{\text{е}} \eta_{k+2}^{\text{ц}} (x_k^{\text{в}} \eta_k^{\text{в}} \vee x_k^{\text{д}} \eta_k^{\text{д}} \vee$$
$$\left. \vee x_k^{\text{н}} \eta_k^{\text{н}} \vee x_k^{\text{р}} \eta_k^{\text{р}} \vee x_k^{\text{т}} \eta_k^{\text{т}})) \vee x_k^{\text{ь}} \eta_k^{\text{е}} \eta_{k+1}^{\text{ц}} \eta_{k+2}^{\sqcup}) \right]. \tag{35}$$

Произведение $L_{5,1} \wedge L_{5,2}$ даст предикат $L_5(\beta, X, \eta, \gamma)$:

$$L_5(\beta, X, \eta, \gamma) = L_{5,1} \wedge L_{5,2}. \tag{36}$$

Последним звеном нашей модели будет описание зависимости между типом склонения γ, окончанием словоформы ω, чередованием β, основой α и словом X. Если тип склонения 6-й, а окончание словоформы -й или же тип склонения 5-й, окончание пус-

тое, а третьей от конца буквой слова X является *н*, *р*, *т* или *ь* (слово *солнце* — исключение), то в словоформе должна иметь место вторая ступень чередования (в противном случае ступень чередования первая):

$$L_6(\alpha,\ X,\ \omega,\ \gamma,\ \beta) = L_{6,1}L_{6,2} \sim \beta^2; \tag{37}$$

$$L_{6,1} = \gamma^6\omega^{\text{й}} \vee \gamma^5\omega^{\sqcup} (\mu_1^{\text{в}} \vee \mu_1^{\text{д}} \vee \mu_1^{\text{н}} \vee \mu_1^{\text{р}} \vee \mu_1^{\text{т}} \vee \mu_1^{\text{ь}})\,\bar{\alpha}^o. \tag{38}$$

Здесь показатель c в выражении $\bar{\alpha}^c$ означает номер основы слова *солнце*, μ_1 — третья от конца буква слова X:

$$L_{6,2} = (\forall i \in \{4,\ 5,\ \ldots,\ n\})\,[x_{i-1}^{\text{е}}x_i^{\sqcup} \supset (\mu_1 = x_{i-3})]. \tag{39}$$

Построенную модель фрагмента морфологии русского языка укрупненно можно представить в виде системы предикатов

$$L_0(X,\ Y,\ Z,\ \Gamma);\ L_1(Z,\ \gamma,\ \omega);\ L_2(X,\ \alpha);\ L_3(X,\ \alpha,\ \gamma);\ L_4,\ (\eta,\ \omega,\ Y);$$
$$L_5(X,\ \beta,\ \eta);\ L_6(\gamma,\ \omega,\ \alpha,\ X,\ \beta).$$

Под предикатом $L_0(X,\ Y,\ Z,\ \Gamma)$ подразумевается конъюнкция предикатов (1—9).

При решении задач морфологической обработки не обязательно представлять систему (40) в виде единого уравнения $L_0 \wedge L_1 \wedge \ldots \wedge L_6 = 1$. Можно отдельно искать корни каждого из уравнений $L_i = 1$ $(i = 0,\ 1,\ \ldots,\ 6)$ и в качестве ответа задачи брать те из них, которые удовлетворяют всем уравнениям системы. При этом, решая очередное уравнение, целесообразно опираться на сведения о корнях уравнений, решенных ранее.

Усилия, затраченные на отыскание корней системы существенно зависят от того, в каком порядке решаются уравнения $L_i = 1$. По-видимому, для каждого типа морфологических задач существует некоторый оптимальный порядок. На наш взгляд, для задачи синтеза он таков: L_0, L_2, L_3, L_1, L_6, L_5, L_4.

Приведем пример морфологического синтеза. Заданы словарная форма *поле* ($x_1 = n$, $x_2 = o$, $x_3 = л$, $x_4 = e$, $x_5 = x_6 = x_7 = \sqcup$), число искомой словоформы — единственное ($z_1 = $ е), падеж — творительный ($z_2 = $ т). Требуется определить словоформу, т. е. найти значение переменной Y. В разбираемых примерах ограничимся величиной $n = 7$. Разделим весь процесс отыскания переменной Y на этапы, каждый из которых будем связывать с решением одного из уравнений $L_i = 1$. Для ясности начало описания очередного этапа обозначим символом соответствующего предиката.

L_0. В результате «решения» уравнения $L_0 = 1$ убеждаемся, что исходные значения X и Z не выходят за пределы их областей определения.

L_2. Значение $X = \langle n,\ o,\ л,\ e,\ \sqcup,\ \sqcup,\ \sqcup,\ \sqcup\rangle$ обращает в 0 все конъюнкции $x_1^{a_{i1}}x_2^{a_{i2}}$, $\ldots$, $x_n^{a_{in}}$, кроме одной — $x_1^n x_2^o x_3^л x_4^e \times$ $\times\, x_5^{\sqcup} x_6^{\sqcup} x_7^{\sqcup}$. Поэтому существует лишь одно значение α, которое

удовлетворяет уравнению $L_2(X, \alpha) = 1$, а именно то, которое обращает в 1 эквивалентность $\alpha^m \sim x_1^n x_2^o x_3^{\text{л}} x_4^e x_5^{\sqcup} x_6^{\sqcup} x_7^{\sqcup}$. Поскольку предикат (23) записан в общем виде, будем считать, что $\alpha = m$, где m — условный номер основы слова *поле*.

L_3. Так как слово *поле* не относится к числу несклоняемых, M_0 не может содержать элемент m (из-за того, что M_0 не задано явно, мы лишены возможности проверить это непосредственно). Значит, справедливо утверждение $\overline{\alpha \in M_0}$, что влечет истинность всех сомножителей произведения $L_{31}L_{32}L_{33}$. Определим значение переменной μ из (26); $x_{i+1}^e x_{i+2}^{\sqcup} = 1$ только, если $i = 3$, значит, $\mu = x_3 = л$. Это согласуется с (25) и в силу (24) дает единственно возможное значение для типа склонения: $\gamma = 2$.

L_1. Так как $\gamma = 2$, из (10) заключаем, что $L(2, z_1, z_2, \omega) = 1$. Из условия $z_1 = $ м и (11) получаем, что $L(2, м, z_2, \omega) = 1$. Наконец, из выражения для $L(2, м, z_2, \omega)$ (13) и условия $z_2 = $ т имеем $\omega =$ *-ями*.

L_6. Значение переменной $\gamma = 2$ обращает в 0 предикат (38). Следовательно, из (37) имеем $0 \sim \beta^2$, что в сочетании с (7) дает $\beta = 1$.

L_5. Выражение (36) говорит о том, что корни системы должны обратить в 1 предикаты $L_{5,1}$ и $L_{5,2}$. Из того, что $\beta = 1$ и $L_{5,1} = 1$, однозначно следуют равенства $\eta_1 = x_1 = п$; $\eta_2 = x_2 = о$; $\eta_3 = x_3 = л$; $\eta_4 = \eta_5 = \eta_6 = \eta_7 = \sqcup$.

L_4. Из (33) и (30) получаем искомое значение переменной $Y = \langle y_1, y_2, \ldots, y_n \rangle$: $y_1 = \eta_1 = п$; $y_2 = \eta_2 = о$; $y_3 = \eta_3 = л$; $y_4 = я$; $y_5 = м$; $y_6 = и$; $y_7 = \sqcup$.

Рассмотрим пример морфологического анализа. Дана словоформа *донец* ($y_1 = д$, $y_2 = о$, $y_3 = н$, $y_4 = е$, $y_5 = ц$, $y_6 = \sqcup$, $y_7 = \sqcup$). Требуется отыскать значение числа и падежа, соответствующие заданной словоформе. Порядок решения уравнений в задаче анализа выберем следующим: L_0, L_4, L_5, L_2, L_3, L_6, L_1.

L_0. Согласно (2), заданные значения переменных y_i ($i = 1, 2, \ldots, 7$) содержатся в их области определения.

L_4. Словоформа *донец* такова, что обращает в 0 выражения, стоящие в предикатах $L_{4,i}$ после знака импликации. Лишь в предикате $L_{4,\sqcup}$ (32) при условии $\eta_i = y_i$ ($i = 1, 2, \ldots, 7$) это выражение может быть равным 1, что допускает значение $\omega = \sqcup$. Это единственно возможное решение, при котором истинны все предикаты $L_{4,i}$ ($i = 1, 2, \ldots, 18$), а вместе с ними и L_4 (33).

L_5. Анализируя предикаты, составляющие L_5 (34)—(36), убеждаемся, что уравнению $L_5 = 1$ удовлетворяют два набора значений неизвестных X и β: $\langle\langle д, о, н, е, ц, \sqcup\rangle, 1\rangle$ и $\langle\langle д, о, н, ц, е, \sqcup, \sqcup\rangle, 2\rangle$.

L_2. Решая уравнение $L_2 = 1$ (23), приходим к выводу, что $X = \langle д, о, н, е, ц, е, \sqcup\rangle$ не является его корнем, так как наборы $\langle a_{i1}, a_{i2}, \ldots, a_{in}\rangle$ представляют собой русские слова,

дополненные пробелами до стандартной длины, и среди них не может быть слова *донеце*. Определяем номер основы слова *донце* (условно принимаем $\alpha = d$).

L_3. Сопоставляя d и M_0, видим, что $d \in M_0$, значит, истинен предикат $\alpha \in M_0$, поскольку $\alpha = d$. Из (28) заключаем, что $L_{3,1}L_{3,2}L_{3,3} = 1$, далее, согласно (26) $\mu = ц$. Это удовлетворяет (25) и в соответствии с (24) дает $\gamma = 5$.

L_6. Подстановка ранее найденных значений $\alpha = d$, $X = <д, о, н, ц, е, \sqcup, \sqcup>$, $\omega = \sqcup$, $\gamma = 5$, $\beta = 2$ в уравнения $L_{6,1} = 1$; $L_{6,2} = 1$; $L_6 = 1$ (37)—(39) не приводит к противоречию, что позволяет нам перейти к следующему этапу анализа.

L_1. Из (10) и $\gamma = 5$ делаем заключение об истинности предиката $L_1(5, z_1, z_2, \omega)$. Анализируя (17) и (18), замечаем, что $L_1(5, е, z_2, \omega)$ ложен при имеющемся значении $\omega = \sqcup$, а $L_1(5, м, z_2, \omega) = 1$ в случае $z_2 = р$ или $z_2 = в$. Из (11) в этих же случаях получаем: $z_1 = м$.

Итак, решив задачу морфологического анализа, мы установили, что число словоформы *донец* — множественное, падеж — родительный или винительный. Это не должно вызывать удивления, если вспомнить, что построенная модель не учитывает категории одушевленности. Полученный ответ допускает оба возможных значения этой категории. Попутно выяснили, что $X = <д, о, н, ц, е, \sqcup, \sqcup>$, т. е. решили задачу нормализации.

Может показаться, что решение некоторых уравнений, например $L_2 = 1$, излишне при морфологическом анализе. Следующий пример доказывает обратное. Пусть требуется произвести морфологический анализ ошибочно заданной словоформы *маре*.

L_4. Уравнению $L_4 = 1$ удовлетворяют две пары значений: $\omega = \sqcup$, $\eta = <м, а, р, е, \sqcup, \sqcup, \sqcup>$ и $\omega = е$, $\eta = <м, а, р, \sqcup, \sqcup, \sqcup, \sqcup>$.

L_5. Возможными корнями уравнения $L_5 = 1$ будут: $X = <м, а, р, е, \sqcup, \sqcup, \sqcup>$, $X = <м, а, р, е, е, \sqcup, \sqcup>$, $X = <м, а, р, \sqcup, \sqcup, \sqcup, \sqcup>$.

L_2. Ни одно из буквосочетаний *маре*, *марее*, *мар* не является словом русского языка, поэтому ни одно из соответствующих значений переменной X не обратит в тождества одновременно уравнения $L_2 = 1$ (23) и (5). Следовательно, не существует решения системы L' в целом, что свидетельствует о противоречивости исходных данных.

Если требуется по заданной словоформе найти ее словарную форму, т. е. произвести нормализацию, необходимо последовательно решить уравнения: $L_0 = 1$; $L_4 = 1$; $L_5 = 1$; $L_2 = 1$. Например, задана словоформа *чудищ* ($y_1 = ч$, $y_2 = у$, $y_3 = д$, $y_4 = и$, $y_5 = щ$, $y_6 = y_7 = \sqcup$).

L_0. С помощью (2) проверяем, принадлежат ли значения переменных y_i ($i = 1, 2, \ldots, 7$) их области определения.

L_4. Из (33), (32) и (6) заключаем, что $\omega = \sqcup$, $\eta = <ч, у д, и, щ, \sqcup, \sqcup>$.

L_5. Уравнение $L_5 = 1$ дает два допустимых значения переменной X: <ч, у, д, и, щ, е, ␣> и <ч, у, д, и, щ, ␣, ␣>.

L_2. Из этих значений уравнению $L_2 = 1$ удовлетворяет лишь первое, т. е. искомая словарная форма — *чудище*.

В заключение отметим, что построенная нами модель данного фрагмента морфологии не является единственно возможной. Состав предикатов этой модели может быть другим, или же они могут выражаться иными формулами.