Е. А. СОЛОВЬЕВА

ОБЩИЙ КЛАССИФИКАТОР ЛИЧНЫХ ФОРМ С УЧЕТОМ ОМОГРАФИИ

Целью настоящей статьи является математическое описание функции f_Z, которую грамотный человек (руководствующийся правилами грамматики [1, 2]) реализует в процессе морфологической классификации личных глагольных форм русского языка по признакам всех собственно грамматических категорий. Так, f_Z (читаю) $= z_1$, f_Z (сидишь) $= z_2$ и т. д., где z_i $(i = 1{,}17)$ характеризуют классы эквивалентности (формальные классы S_Z личных форм по признаку $Z = \langle M, T, N, L, G \rangle$, т. е. по признакам M (признак категории наклонения), T (времени), N (числа), L (лица) и G (рода) одновременно.

Функция f_Z, как и многие другие, — проявление универсальной морфологической функции грамматической обработки слов. Описание f_Z вносит вклад в описание морфологической функции.

Постановка задач сформулирована ранее [3, 4], функция $f_Z = \langle F_2, X, Y_Z \rangle$, названная общим классификатором личных форм, введена в работе [3] (индекс f опускаем). Область определения X составляют синтетические личные формы глаголов русского языка; область значений $Y_Z = \{z_1, ..., z_{17}\}$ — признаки классов эквивалентности в разбиении $\sigma_Z = \{S_{Z_1}, ..., S_{Z_{17}}\}$ множества X по признаку Z.

Выразим формальные классы $S_{Z_1} - S_{Z_{17}}$ через классы эквивалентности в разбиениях σ_M, σ_T, σ_N, σ_L и σ_G множества X по каждому из рассматриваемых признаков в отдельности:

$$S_{Z_1} = S_{M_1} \cap S_{T_1} \cap S_{N_1} \cap S_{L_1} \cap S_{G_4},$$
$$S_{Z_2} = S_{M_1} \cap S_{T_1} \cap S_{N_1} \cap S_{L_2} \cap S_{G_4},$$
$$S_{Z_3} = S_{M_1} \cap S_{T_1} \cap S_{N_1} \cap S_{L_3} \cap S_{G_4},$$
$$S_{Z_4} = S_{M_1} \cap S_{T_1} \cap S_{N_2} \cap S_{L_1} \cap S_{G_4},$$
$$S_{Z_5} = S_{M_1} \cap S_{T_1} \cap S_{N_2} \cap S_{L_2} \cap S_{G_4},$$
$$S_{Z_6} = S_{M_1} \cap S_{T_1} \cap S_{N_2} \cap S_{L_3} \cap S_{G_4},$$
$$S_{Z_7} = S_{M_1} \cap S_{T_2} \cap S_{N_1} \cap S_{L_4} \cap S_{G_1},$$

$$S_{Z_8} = S_{M_1} \cap S_{T_2} \cap S_{N_1} \cap S_{L_4} \cap S_{G_2},$$
$$S_{Z_9} = S_{M_1} \cap S_{T_2} \cap S_{N_1} \cap S_{L_4} \cap S_{G_3},$$
$$S_{Z_{10}} = S_{M_1} \cap S_{T_2} \cap S_{N_2} \cap S_{L_4} \cap S_{G_4},$$
$$S_{Z_{11}} = S_{M_2} \cap S_{T_2} \cap S_{N_1} \cap S_{L_2} \cap S_{G_4},$$
$$S_{Z_{12}} = S_{M_2} \cap S_{T_3} \cap S_{N_2} \cap S_{L_1} \cap S_{G_4},$$
$$S_{Z_{13}} = S_{M_2} \cap S_{T_3} \cap S_{N_2} \cap S_{L_2} \cap S_{G_4},$$
$$S_{Z_{14}} = S_{M_2} \cap S_{T_4} \cap S_{N_1} \cap S_{L_2} \cap S_{G_4},$$
$$S_{Z_{15}} = S_{M_3} \cap S_{T_4} \cap S_{N_1} \cap S_{L_5} \cap S_{G_4},$$
$$S_{Z_{16}} = S_{M_3} \cap S_{T_4} \cap S_{N_2} \cap S_{L_2} \cap S_{G_4},$$
$$S_{Z_{17}} = S_{M_3} \cap S_{T_5} \cap S_{N_3} \cap S_{L_6} \cap S_{G_4}.$$

Формальные категории $\sigma_M = \{S_{M_1}, S_{M_2}, S_{M_3}\}$ $\sigma_T = \{S_{T_1}, \ldots S_{T_5}\}$, $\sigma_N = \{S_{N_1}, S_{N_2}, S_{N_3}\}$ и $\sigma_L = \{S_{L_1}, \ldots S_{L_6}\}$ эксплицируют соответственно грамматические категории

$$\sigma'_M = \{S'_{M_1}, S'_{M_2}\}, \quad \sigma'_T = \{S'_{T_1}, S'_{T_2}, S'_{T_3}\}, \quad \sigma'_N = \{S'_{N_1}, S'_{N_2}\}$$
и $\sigma'_L = \{S'_{L_1}, \ldots S'_{L_4}\}$ [5], где $S_{M_1} = S'_{M_1} \setminus S'_{M_2}$, $S_{M_2} =$
$= S'_{M_2} \setminus S'_{M_1}$, $S_{M_3} = S'_{M_1} \cap S'_{M_2}$, $S_{T_1} = S_{T_1} \setminus S'_{T_3}$, $S_{T_2} = S'_{T_2} \setminus S'_{T_3}$,
$S_{T_3} = S'_{T_3} \setminus (S'_{T_1} \cup S'_{T_2})$, $S_{T_4} = S'_{T_1} \cap S'_{T_3}$, $S_{T_5} = S'_{T_2} \cap S'_{T_3}$, $S_{N_1} =$
$= S'_{N_1} \setminus S'_{N_2}$, $S_{N_2} = S'_{N_2} \setminus S'_{N_1}$, $S_{N_3} = S'_{N_1} \cap S'_{N_2}$, $S_{L_1} = S'_{L_1} \setminus S'_{L_2}$, $S_{L_2} =$
$= S'_{L_2} \setminus (S'_{L_1} \cup S'_{L_4})$, $S_{L_3} = S'_{L_3}$, $S_{L_4} = S'_{L_4} \setminus S'_{L_2}$, $S_{L_5} = S'_{L_1} \cap S'_{L_2}$,
$S_{L_6} = S'_{L_2} \cap S'_{L_4}$.

Категория рода $\sigma_G = \sigma'_G = \{S_{G_1}, \ldots, S_{G_4}\}$. Признаки грамматических значений описаны в табл. 1.

График $F_Z \subset X \times Y_2$ функции f_Z можно представить в виде [6]

$$F_Z = \varepsilon \{\langle x, y \rangle | K_Z(x, y)\},$$

где $K_Z(x, y)$ — двухместная высказывательная форма. Если $\langle x, y \rangle \in F_Z$, то при подстановке такой пары в $K_Z(x, y)$ получим истинное высказывание. Удобный способ описания $K_2(x, y)$ разработан нами ранее [3, 5].

Переобозначим выражения, наиболее часто встречающиеся в записи $K(x, y)$: $P(x) \cap P_{A_i} \neq \varnothing$ обозначим через $A_i(x)$; $P(x) \cap P_{A_i} = \varnothing$ — через $\bar{A}_i(x)$; $P_d(x) \cap P_{D_j} \neq \varnothing$ — через $D_j(x)$, $P_d(x)$

Таблица 1

Признак	Грамматический смысл признака
M_1'	Изъявительное наклонение
M_2'	Повелительное наклонение
T_1'	Непрошедшее время
T_2'	Прошедшее время
T_3'	Нет времени
N_1'	Единственное число
N_2'	Множественное число
L_1'	1-е лицо
L_2'	2-лицо
L_3'	3-лицо
L_4'	Нет лица
G_1'	Мужской род
G_2'	Женский род
G_3'	Средний род
G_4'	Нет рода

$\cap P_{D_j} = \varnothing$ — через $\bar{D}_j(x)$; $y = y_k$ — через $R_k(x)$. Для дальнейшего сокращения записи заменим $A_i(x)$, $\bar{A}_i(x)$, $D_j(x)$, $\bar{D}_j(x)$, $R_k(y)$ на A_i, $\bar{A}_i$, D_j, $\bar{D}_j$, R_k соответственно. Состав множеств формальных признаков P_{A_i} и P_{D_j}, который при необходимости может быть ограничен или дополнен без изменения структуры $K_2(x, y)$, определяется в результате исследований или использования имеющихся данных грамматики и работ в области автоматического анализа.

Грамматика [1, 2 и др.] не содержит данных о том, как формально различать омографию словоформ, и многих других нужных сведений. Поэтому для построения алгоритмов автоматического анализа «сначала должна быть проделана весьма трудоемкая лингвистическая работа по формализованному описанию языков» [7, с. 12]. В известных литературных источниках также нет сведений по разбору омографии, но указывается, например, в [8], на важность (а также сложность) получения таких данных. Поэтому был произведен автоматический разбор всех случаев омографии личных глагольных форм [9], необходимый для описания $K_Z(x, y)$.

В результате проведенных исследований процессов морфологической классификации установлено, что

$$K_Z(x, y) = [\bigwedge_{i=1}^{i=13} (A_i \Longrightarrow R_i) \wedge (A_{14} \wedge \bar{A}_{15} \Longrightarrow R_7) \wedge (A_{16} \wedge \bar{A}_{10} \Longrightarrow$$

Таблица 2

i	P_{A_i}
1	2
1	у, ю, дам
2	ишь
3	есть, ет, ит, ст
4	им
5	ете
6	суть, ут, ют, ат, ят
7	л, з, с, б, п, р, к, х
8	ла
9	ло
10	скорбели
11	$P_{A_{17}} \cup P_{A_{18}}$
12	ем, им, мте, ем-ка, им-ка, ите-ка
13	йте, ьте, йте-ка, ьте-ка, ите-ка
14	г
15	ляг
16	бели
17	барахли, боли, весели, вызволи, вызоли, выпяли, вытрали, дли, кабали, кайли, коли, кругли, моли, поли, пыли, рули, сверли, светли, скобли, скули, соли, стрели, сули, сусли, тепли, тяжели, утоли, холи, цели, числи, определи, стели
18	й, й-ка, ь-ка, би, ви, ги, ди, жи, зи, ки, й-ка, ми, ни, пи, ри, си, ти, фи, чи, ши, щи, завись, обезопась, гундось, тулумбась, чудесь, торось, весь, высь, брось, квась, крась, ляг
19	ли
20	ешь
21	усь
22	ем
23	ите

Таблица 3

j	P_{D_j}
1	2
1	ем, въемся, взъемся, выем, доем, заем, изъем, наем, надоем, недоем, объем, отъем, переем, подъем, поем, приемся, проем, разъем, съем, уем
2	вали, ввали, взвали, вывали, завали, навали, обвали, отвали, перевали, привали, повали, подвали, провали, развали, свали, ували, открыли, достекли, застекли, недостекли, остекли, перестекли, окисли, выдели, додели, недодели, обдели, отдели, передели, подели, удели, измели, обмелись, перемелись, размелись, смелись, воспались, опались, распали, пились, выпились, напили, надпили, распились, шлись, разошли, ниспошли, телись, отелись, выкали, докали, закали, накали, обкали, перекали, подкали, прокали, раскали, дошались, зашали, нашали, пошали, расшались, умали
3	струсь, перетрусь
4	потешь, тешь, утешь
5	$\bigcup_{i=2}^{i=4} P_{D_i}$
6	пишите, тяните, ...
7	ешь, въешься, взъешься, выешь, доешь, заешь, изъешь, наешь, надоешь, недоешь, объешь, отъешь, переешь, подъешь, поешь, приешься, проешь, разъешь, съешь, уешь
8	трусь, вытрусь
9	сидите, ходите, ...
10	вымыли, вели, повели, стекли, кисли, перекисли, раскисли, дели, надели, одели, раздели, вымели, домели, замели, намели, обмели, отмели, перемели, примели, подмели, помели, промели, размели, смели, пали, выпали, запали, опали, перепали, припали, подпали, попали, пропали, распались, спали, пили, выпили, допили, запили, испили, опили, отпили, перепили, подпили, попили, пропили, распили, шли, вышли, дошли, зашли, нашли, отошли, перешли, пришли, пошли, подошли, сошли, ушли

$$\Rightarrow R_{11}) \wedge (D_1 \Rightarrow R_1) \wedge (D_5 \Rightarrow R_{11}) \wedge (\bigwedge_{j=6}^{j=10} D_j \Rightarrow R_{j+7}) \wedge (A_{19} \wedge \bar{A}_{17}$$
$$\wedge \bar{D}_2 \wedge \bar{D}_{10} \Rightarrow R_{10}) \wedge (A_{20} \wedge \bar{D}_4 \wedge \bar{D}_7 \Rightarrow R_2) \wedge (A_{21} \wedge \bar{D}_3$$
$$\wedge \bar{D}_8 \Rightarrow R_1) \wedge (A_{22} \wedge \bar{D}_1 \Rightarrow R_4) \wedge (A_{23} \wedge \bar{D}_6 \wedge \bar{D}_9 \Rightarrow R_5).$$

Множества формальных признаков P_{A_i} $(i = \overline{1,23})$ и P_{D_j} $(j = \overline{1,10})$ приведены в табл. 2 и 3 соответственно.

Знак $\sim$ введен для упрощения записи. Он означает, что имеется в виду не только приведенный в таблице признак, но и полученный из него приписыванием подстфикса *ся* (*сь*) за буквой, отмеченной сверху волнистой чертой. Например, $P_{A_1} = \{$ у, ю, ются, дам, дамся $\}$. Множества P_{D_6} и P_{D_9} не даны здесь полностью, так как число их элементов сравнительно велико. Большая часть глаголов из P_{D_6} и P_{D_9} в рассматриваемых формах малоупотребительна, поэтому в модификации f_Z^4 предложенной модели эти множества составлены на основании частотного словаря [10]. Знак ударения, позволяющий в ряде случаев (например, в некоторых формах на *ли*) упростить классификацию, не использовался, так как в реальных текстах он обычно отсутствует.

При подстановке пары ⟨читаю, z_1⟩ в форму $K_z(x, y)$ получим истинное высказывание, т. е. K_Z (читаю, z_1) = 1, значит ⟨читаю, z_1⟩ $\in F_Z$. Кроме того, пары ⟨читаю, z_i⟩ при $i = \overline{2,17}$ обратят $K_2(x, y)$ в 0, следовательно, f_Z (сидишь) = z_1. Аналогично определим, что f_Z (читаем) = z_2, f_Z (ходим) = z_{10}, f_Z (вели) = = z_{17} и т. д.

Общий классификатор f_Z строился таким образом, чтобы безошибочно обрабатывать синтетические личные формы глаголов из словаря [10] на 104 тыс. слов. Наибольшие трудности при моделировании рассматриваемого процесса классификации связаны с автоматическим разбором полного и частичного совпадения словоформ, особенно форм на *ли*. Для разрешения омографии проведены эксперименты с человеком на словах и псевдословах [5], тщательный формальный анализ форм глаголов из словаря [11].

Отмегим лишь словоформу *поем* (*поем* от *поесть* и *поём* от *петь*), которая будет омографичной, если не различать буквы *е* и *ё*, как это делается в реальных текстах. При различении *е* и *ё* в множество P_{A_3} необходимо включить *ёт*, а в P_{A_4} — *ём* (при этом формы *поем* и *поём* будут классифицироваться верно), в противном случае для безошибочного действия модели на форме поем необходимо ввести для этой формы класс $S_{M_{18}} = S_{M_1} \cap \cap S_{T_1} \cap S_{N_3} \cap S_{L_1} \cap S_{G_4}$ и соответствующим образом изменить $K_Z(x, y)$. При практическом действии модели вопрос может быть решен в зависимости от конкретных требований.

Таким образом, на основании предложенной методики разработана универсальная модель f_Z, учитывающая все случаи омографии [8] и использующая минимально необходимую информацию о формальной структуре слова. Предложен ряд модификаций модели f_Z, в частности простые и универсальные модели [3] классификации личных форм по признаку Z без учета омографии. Анализаторы (модели, не учитывающие омографию) описывают функции, приближенные к тем, которые человек реализует на уровне контекста (словоформе ставятся в соответствие наборы признаков из табл. 1). Эти модели обладают высокими вероятностными оценками [3 и др.], так как основаны на данных статистического анализа текстов.

Классификатор f_Z представлен в виде алгоритма [3] на языке граф-схем, все разработанные модели (частные и общие) и их модификации также описаны в виде алгоритмов. Алгоритмы реализованы на ЭЦВМ «Урал-14Д» с помощью транслятора АЛГОЛ-ЦЭМИ, удобного для обработки словоформ. Одна из модификаций модели f_Z и ряд других моделей приняты в Республиканский фонд алгоритмов и программ. Доведение моделей до уровня действующих позволяет использовать их на практике.

Общий классификатор личных форм вошел в комплекс алгоритмов и программ для подсистемы автоматического анализа автоматизированной системы лингвистической обработки документов. Модель f_Z может представить и самостоятельный интерес, как математическое описание некоторой психической функции человека.

СПИСОК ЛИТЕРАТУРЫ

1. Грамматика современного русского литературного языка. Отв. ред. Шведова Н. Ю. М., «Наука», 1970. 767 с.
2. Грамматика русского языка. Т. I. Ред. коллегия: акад. Виноградов В. В. и др. М., Изд-во АН СССР, 1960. 719 с.
3. Соловьева Е. А. Моделирование процессов морфологической классификации с учетом омографии. — В кн.: Проблемы бионики. Вып. 17, Харьков, 1976, с. 126—134.
4. Соловьева Е. А. Анализатор и классификатор личных форм по наклонению и их исследование. — В кн.: Проблемы бионики. Вып. 17, Харьков, 1976, с. 118—126.
5. Соловьева Е. А. К вопросу о построении общего алгоритма морфологической классификации глагольных форм русского языка. — В кн.: Проблемы бионики. Вып. 15, Харьков, 1975, с. 143—149.
6. Шиханович Ю. А. Введение в современную математику. М., «Наука», 1965. 376 с.
7. Кулагина О. С., Мельчук И. А. Автоматический перевод: краткая история, современное состояние, возможные перспективы. — В кн.: Автоматический перевод. М., «Прогресс», 1971, с. 395.
8. Мельчук И. А. Морфологический анализ при машинном переводе (преимущественно на материале русского языка). — В кн.: Проблемы кибернетики. Вып. 6, М., Физматгиз, 1961, с 207—276.
9. Соловьева Е. А. Исследование процессов классификации омографичных глагольных форм. — В кн.: Проблемы бионики. Вып. 16, Харьков, 1976, с. 104—114.