10. Штейфельдт Э. А. Частотный словарь современного русского литературного языка. Таллин, 1963. 316 с.
11. Орфографический словарь русского языка (около 104 тыс. слов). Изд. Редактор Бархударов Ф. Г. и др. М., «Сов. энциклопедия», 1971. 5

Поступила 20 мая 19

УДК 62.506.2

А. И. ЧУГУН

ОБ ОДНОМ МЕТОДЕ КЛАССИФИКАЦИИ ГЛАГОЛЬНЫХ ФОРМ РУССКОГО ЯЗЫКА ПО ТИПАМ СПРЯЖЕНИЯ

Моделирование процессов образования различных форм одного и того же слова по его словарной форме является одним из важнейших этапов исследования способностей человека решать задачи морфологической классификации. Как правило, справочные пособия, за исключением случаев особо сложных образований, указывают только словарную (нормализованную) форму слова той или иной части речи, поэтому моделирование процессов формообразования словоформ по их нормализованным представителям представляет и практический интерес.

Для грамотного человека задачи подобного класса не вызывают затруднения и решаются им на основании опыта, накопленного при обучении речи, и знаний неформальных правил традиционной грамматики. Опыт моделирования речевого поведения человека при решении задач морфологической классификации [1, 2] показывает, что автоматический анализ и синтез морфологических цепочек требуют формализованного подхода к их решению, т. е. создание действующей модели решения задачи формообразования слов предполагает описание функций, реализуемых человеком, в виде системы формальных правил.

Анализируя словарные формы любой части речи, нетрудно заметить, что решение задачи формообразования (словоизменения) подразделяется на два последовательных этапа: 1-й этап — распределение нормализованных словоформ по типам словоизменения; 2-й этап — определение системы формальных правил собственно процесса словоизменения. Целью данной работы является выявление системы формальных правил (признаков) процесса распределения словарных форм по типам словоизменения на примере инфинитива невозвратных глаголов русского языка и построение алгоритма их автоматической классификации, эквивалентного этой системе.

Имеется множество A, включающее в себя все невозвратные глаголы русского языка в форме инфинитива. Учитывая динамический характер языка (на время решения поставленной задачи), зафиксируем его развитие с помощью словаря [3]. Тогда можно характеризовать множество A как конечное множество, мощность которого определяется количеством элементов, входящим в него [4] (количеством словоформ, зафиксированным в словаре).

Каждый элемент этого множества $x_i \in A$ представляет собой одну словоформу и $\bigcup_{i=1}^{n} x_i = A$ (n — мощность множества A). Из грамматики [5] известно, что в систему словоизменения (спряжения) глагола входят изменения по наклонениям, временам, числам, лицам и родам. Спрягая элементы x_i по этим грамматическим категориям, получим конечное множество B — множество словоформ личных форм глаголов (здесь и далее под термином «глагол» подразумеваются невозвратные глаголы). Элементы $x_j \in B$ ($\bigcup_{j=1}^{m} x_j = B$) представляют собой так же отдельные словоформы, т. е. во множество B не входят частицы *бы*, *пусть*, *да* и т. п.

Используя эти свойства элементов x_j и принцип сочетаемости грамматических категорий для глаголов, известный из грамматики, априори можно заявить, что каждый элемент $x_i \in A$ при спряжении будет порождать 12 элементов $x_j \in B$. Объединение множеств $A \cup B = M$ — определяет множество всех глаголов русского языка в форме инфинитива и личных спрягаемых форм, $M = \{ x : x \in A$ или $x \in B \}$ и $\bigcup_{k=1}^{l} x_k = M$ ($l = n + m$ — мощность множества M). Элементы $x_k \in M$ можно объединить по признаку лексического значения. Тогда данное множество разобьется на подмножества N_i ($\bigcup_{i=1}^{n} N_i = M$), представляющие собой отдельные парадигмы соответствующих основ глаголов.

Отсюда можно заметить, что $M \sim A$, так как между элементами этих множеств существует взаимно однозначное соответствие ($x_i \leftrightarrow N_i$). Каждое подмножество N_i состоит из 13 элементов $x_k \in N_i$ ($N_i \subset M$).

В процессе анализа подмножеств N_i было получено, что элементы $x_k \in N_i$ имеют один общий отличительный признак, характерный для данного подмножества, и индивидуальные признаки, присущие каждому отдельному элементу этого же подмножества. Общий признак назовем базой парадигмы, а индивидуальные — системой окончаний парадигмы. База парадигмы это часть словоформы, не изменяемая в процессе спряжения, а саму парадигму можно определить как совокупность словоформ, характеризующуюся одной базой и определенной системой окончаний.

Следует отметить, что между понятиями базы парадигмы, системы окончаний парадигмы и основой и окончаниями словоформ при делении ее на морфемы имеется существенное различие. Так, традиционная грамматика для системы глаголов в форме инфинитива и личных спрягаемых форм рассматривают два типа основ (хотя в большинстве случаев они совпадают) — основу инфинитива и прошедшего времени, изъявительного наклоне-

ния и основу личных форм непрошедшего времени и форм повелительного наклонения. Например, для слова *брать* имеются основы *бра-* и *бер-*, а слово *бегать* — имеют одну основу для всех форм — *бега-*, тогда как база парадигмы слова *брать* (*брать, беру, берешь, берет, берем, берете, берут, брал, брала, брало, брали, бери, берите*) будет представлена неизменяемой частью словоформы и состоять из одной буквы *б-*, остальные части словоформ образуют систему окончаний парадигмы (*-рать, -еру, -ерешь, -ерет, -ерем, -ерете, -ерут, -рал, -рала, -рало, -рали, -ери, -ерите*). Такое деление словоформ предполагает, что все элементы множества M относятся только к флективным классам слов [6].

Если база парадигмы служит отличительным признаком подмножеств N_i, то многие из них характеризуются и общим признаком, в качестве которого выступает система окончаний парадигмы. По этому признаку все множество подмножеств N_i разбиваем на классы T_c. Каждый из этих классов будет характеризоваться одной, только ему присущей, системой окончаний парадигмы. Классы будем называть типами словоизменения или, по аналогии с традиционной грамматикой, типами спряжения.

Предварительный анализ элементов $x_i \in A$ показал, что классификацию их по типам спряжения можно проводить, основываясь на формальные признаки, заложенные в самой нормализованной словоформе. Для классификации инфинитива в качестве таких признаков были использованы: окончания словоформ; буква или буквосочетание, стоящие перед окончанием; место ударения в словоформе (ударение падает на основу или на окончание). В результате экспериментов получено, что все множество подмножеств N_i разбивается на 124 типа спряжения. Из них первые 24 включают в себя все продуктивные глаголы, а по остальным распределяются непродуктивные глаголы русского языка. Окончания, которые нужно отбрасывать от инфинитива в соответствующем типе спряжения для выделения из этой формы базы парадигма и номера типов спряжения, представлены в таблице.

Присвоение номеров типов спряжения осуществляется произвольно и зависит только от порядка окончаний, имеющихся в инфинитиве. В нашем случае окончания проверяются последовательно в следующем порядке: *-ать, -ять,-еть, -уть, -оть, -ыть, -сть, -зть, -ить, -чь, -ти* (по продуктивным глаголам).

Алгоритм автоматической классификации невозвратных глаголов по типам спряжения состоит из 11 блоков. Каждый блок настроен на проверку одного из окончаний, приведенных выше, и представляет собой набор процедур, которые в своей работе используют формальные признаки, выявленные при классификации инфинитива. В качестве основных формальных процедур для построения алгоритма выбраны: проверка окончания словоформы на совпадение с эталонным окончанием; проверка буквы перед окончанием; проверка буквосочетания перед окончанием

Тип спряжения	Окончание инфинитива	Тип спряжения	Окончание инфинит.	Тип спряжения	Окончание инфинитива	Тип спряжения	Окончание инфинитива	Тип спряжения	Окончание инфинит.
1	ть	26	вать	51	хать	76	еть	101	чь
2	овать	27	вать	52	тать	77	нуть	102	ечь
3	евать	28	ать	53	тать	78	нуть	103	жечь
4	евать	29	ать	54	тать	79	нуть	104	чь
5	ять	30	ать	55	стать	80	нуть	105	ечь
6	уть	31	ать	56	ть	81	олоть	106	чь
7	уть	32	ать	57	ъять	82	оть	107	чь
8	ить	33	ать	58	ъять	83	ыть	108	чь
9	ить	34	жать	59	ять	84	ть	109	очь
10	ить	35	гать	60	ять	85	ыть	110	чь
11	ить	36	гать	61	мять	86	сть	111	идти
12	ить	37	дать	62	ять	87	сть	112	ти
13	дить	38	ть	63	нять	88	ть	113	ти
14	дить	39	дать	64	нять	89	сть	114	йти
15	зить	40	кать	65	зять	90	есть	115	асти
16	зить	41	кать	66	деть	91	есть	116	сти
17	стить	42	скать	67	еть	92	есть	117	сти
18	ить	43	ать	68	еть	93	сть	118	сти
19	тить	44	рать	69	еть	94	ть	119	бить
20	тить	45	нать	70	теть	95	ть	120	вить
21	сить	46	слать	71	стеть	96	ить	121	пить
22	сить	47	лать	72	теть	97	ить	122	шить
23	ить	48	ать	73	среть	98	лить	123	зать
24	ить	49	сать	74	еть	99	ить	124	ять
25	ать	50	хать	75	сеть	100	тить		

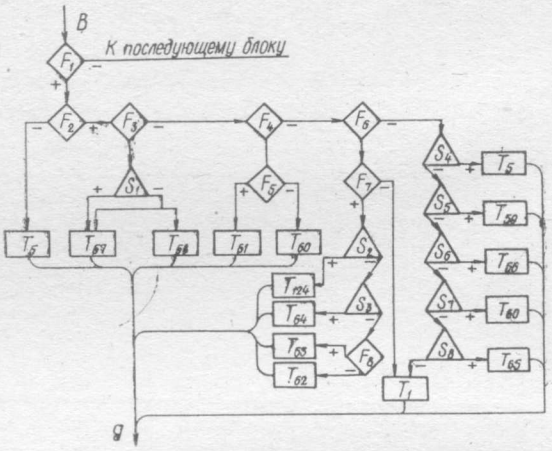
или буквой, стоящей перед окончанием; проверка местоположения ударения в словоформе; присвоение номера типа спряжения; отбрасывание соответствующего окончания.

Проверка работы алгоритма показала, что входное множество A содержит 275 словоформ, обладающих формальными признаками того или иного типа спряжения, но не принадлежащих этим типам спряжения. Эти словоформы составляют $\approx 1\%$ от входного множества, выделены в отдельные словари и дополнительными процедурами для их классификации являются: проверка словоформ, поступающей на вход, на полное совпадение со словарем-эталоном; отбрасывание части словоформы, совпавшей со словарем; проверка оставшейся части словоформы на совпадение с набором приставок. Для алгоритма составлены 97 словарей-эталонов, 57 из которых состоят из одной словоформы, 27 — из 2-х ÷ 5-и, 10 — из 6-и ÷ 10-и и 3 с количеством словоформ в каждом, превышающем 10. Максимальная длина одного из словарей — 25 словоформ. Словари-эталоны распределены соответствующим образом по блокам алгоритма.

Привести в данной статье всю блок-схему алгоритма не представляется возможным из-за ее разветвленности. Поэтому для иллюстрации его работы на рисунке показана блок-схема только

одного из 11 блоков; а именно блока автоматической классификации по типам спряжения глаголов, оканчивающихся в инфинитиве на *-ять*.

Из рисунка видно, что алгоритм состоит из элеметарных блоков — распознавателей и операторов. Распознаватели F_i и S_j — логические блоки, которые проверяют наличие у словоформы, по-



ступающей на вход B того или иного формального признака. Операторы T_k присваивают словоформе номер типа спряжения и согласно этому номеру отбрасывают окончание, помещенное в таблицу, и подают на выход g — номер типа спряжения, базу парадигмы и окончание в инфинитиве.

Распознаватели F_i, в порядке расположения их в блок-схеме, проверяют в словоформе, поступающей на вход, следующие формальные признаки: F_1 — сравнивает окончание словоформы с эталонным окончанием (*-ять*); F_2 — проверяет место ударения в словоформе; F_3 — проверяет первую букву перед окончанием на совпадение с (-ъ-); F_4 — проверяет первую букву перед окончанием на совпадение с (-М-); F_5 — проверяет вторую букву перед окончанием на совпадение с гласными; F_6 — проверяет первую букву перед окончанием на совпадение с (-Н-); F_7 — проверяет часть слова перед буквой (-Н-) на совпадение с приставкой; $F_8 = F_5$, а S_j сравнивает словоформу со словарями-эталонами: S_1 — изъять; S_2 — внять; S_3 — принять; S_4 — осмеять; S_5 — стоять; S_6 — застрять; S_7 — распять; S_8 — взять. Этот блок классифицирует все без исключения глаголы с окончанием *-ять* в инфинитиве. *Словоформы, не содержащие признака инфинитива, подаются со входа на выход с пометкой:* «Не инфинитив невозвратных глаголов».

Решение задачи автоматической классификации глаголов по типам спряжения позволяет выделить базу любой парадигмы по

ее метке (словарной форме) и результаты работы можно применить на практике для построения алгоритма синтеза личных форм глаголов.

СПИСОК ЛИТЕРАТУРЫ

1. 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.
2. Бондаренко М. Ф., Осыка А. Ф. Об одном алгоритме склонения числительных русского языка. — В кн.: Проблемы бионики. Вып. 15, Харьков, 1975, с. 153—158.
3. Орфографический словарь русского языка. Изд. 12-е, М., «Сов. энциклопедия», 1973. 520 с.
4. Кон П. Универсальная алгебра. М., «Мир», 1968. 351 с..
5. Грамматика русского языка. Т. 1, М., Изд-во АН СССР. 1960. 719 с.
6. Белоногов Г. Г., Богатырев В. И. Автоматизированные информационные системы. М., «Сов. радио», 1973. 250 с.

Поступила 10 апреля 1976 г.