УДК 510.62

М. Ф. БОНДАРЕНКО, канд. техн. наук, *В. И. РУБЛИНЕЦКИЙ*

СЛОВО В ВЫЧИСЛИТЕЛЬНОЙ ЛИНГВИСТИКЕ

В последнее время разрабатываются системы, цель которых обеспечить общение человека с вычислительной машиной на естественном языке. Этот круг вопросов постепенно оформляется в самостоятельную научную дисциплину, и ее все чаще называют вычислительной лингвистикой (см., например, [1]). В вычислительной лингвистике заново ставятся многие вопросы лингвистики традиционной. Обычно эти вопросы ставятся в упрощенной форме, но по необходимости требуют точного и конструктивного ответа. В частности, вычислительная лингвистика нуждается в определении слова, ибо системы вычислительной лингвистики (ниже для краткости ВЛ-системы) на основании этого определения выделяют слово из текста.

Нами дано определение слова для ВЛ-систем, работающих с письменной формой русского языка, и на основании этого определения предложен алгоритм выделения слова из текста.

Традиционная лингвистика общепринятого определения слова выработать не смогла. В. А. Звегинцев [2, с. 51] так характеризует ситуацию: «Пожалуй, самая сомнительная и капризная репутация у слова, несмотря на то, что некоторые лингвисты готовы объявить его основной единицей языка. Слово, как говорит Э. Сепир, неизменно присутствует в сознании, и, следовательно, обладает бесспорной психологической реальностью, что свидетельствуется и тем (здесь мы опять обращаемся к Э. Сепиру), что и совершенно неграмотный человек не испытывает никакого затруднения при расчленения речи на слова. Но как

только доходит до определения слова, оно становится неуловимым, как синяя птица».

Исследователи, придерживающиеся машинной, формальной ориентации, например Р. Г. Пиотровский [3], предложили определение слова письменной формы языка в следующем виде: слово — набор букв алфавита, ограниченный с обеих сторон проблемами.

Такое определение не верно. С одной стороны, оно пропускает бессмысленные наборы букв, как *ПВГДРК,* за которым и «совершенно неграмотный человек» не признает статуса слова, а с другой — не пропускает такого бесспорного слова, как *по-прежнему* (кстати, недавно писавшегося вместе), потому что в его состав входит дефис. Определение страдает еще одним пороком: если слова уже выделены и окружены пробелами, то выделение проделано на основании другого определения (какого?), а если слова входят в текст, то там они, помимо пробелов, могут отделяться знаками препинания — слева (например, открывающей скобкой), справа (например, запятой), слева и справа (например, кавычками). Однако рациональное зерно в обсуждаемом определении есть: в нем приведены (хотя и не все) необходимые условия того, чтобы **строка символов** (это строгий термин, определяемый в синтаксисе алгоритмических языков) была словом, как его интуитивно понимают. Мы пополним необходимые условия, а потом сформулируем достаточные.

Определение 1. Строка символов называется *словоподобной,* если и только если она имеет вид $\sigma\alpha_1\ \alpha_2 \ldots\ \alpha_n\sigma$, где σ— знак препинания или пробел, а α_i, $i=1, \ldots, n$, принадлежит множеству, объединяющему алфавит и дефис.

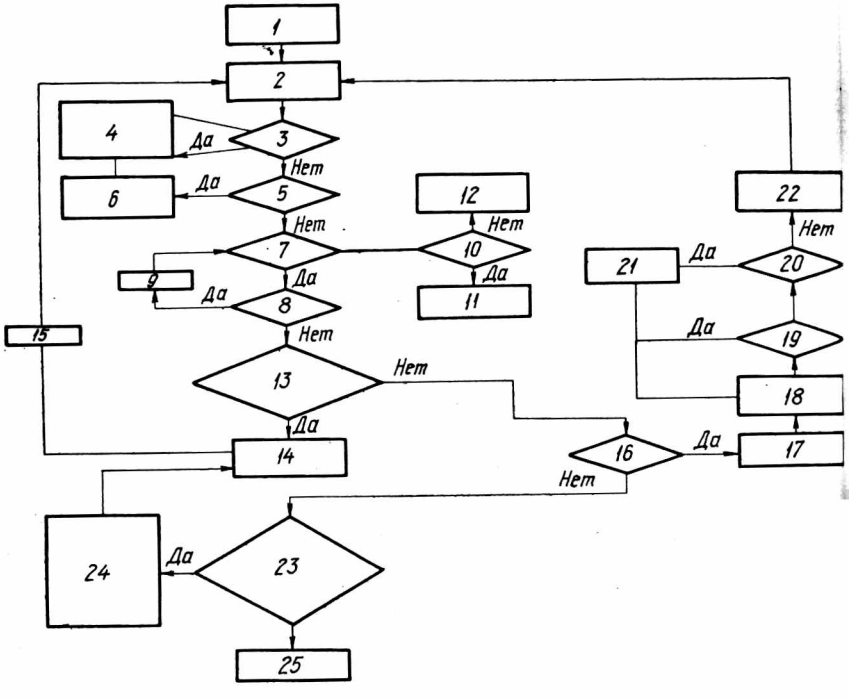
Замечание. Поскольку мы рассматриваем ВЛ-системы, работающие с русским языком, алфавит состоит из всех русских букв. В зависимости от текстов алфавит можно расширять, например, цифрами.

Словоподобная строка символов (ниже для краткости ССС) еще не есть слово, как мы его понимаем интуитивно, ибо, с одной стороны, под определение подпадают по-прежнему бессмысленные строки, как *ПВГДРК,* а с другой стороны, из-за включения дефиса в алфавит, под определение попали словосочетания, такие как осетин-извозчик. Будем называть ССС, которая не является словом в интуитивном смысле, псевдословом. Чтобы сформулировать достаточные условия, обсудим, как отличает слово от псевдослова человек.

Реакцию человека на слово/псевдослово, по-видимому, можно разделить на следующие классы. 1. Слово просто вспоминается. Так, типичный носитель русского языка помнит слово *стол.* Видимо, слово *столами* также не прогоняется по морфологической парадигме, а вспоминается непосредственно. 2. Слово отсутствует в памяти: носитель языка его никогда не слышал и не употреблял. Пожалуй, неологизмы Маяковского «Земля...

заерзает *мясами»* или «В тупой полицейской *слоновости»* являются действительно новыми словами для каждого, кто встретил их первый раз. Но человек владеет словоизменительными и словообразовательными алгоритмами, которые позволяют свести новые слова к известным. 3. Слово неизвестно носителю языка и не приводится к известным через словоизменительные и словообразовательные алгоритмы, однако смысл слова опознается по ситуации. Пример: Балаганов не понял, что означает «статус-кво». Но он ориентировался на интонацию, с которой эти слова были произнесены. 4. ССС неизвестна носителю языка, не преобразуема к известному слову и не угадывается. В этом случае человек не может отличить слово от псевдослова.

То, что является правдоподобной догадкой относительно психологической деятельности человека, в несколько рекомбинированной форме служит твердым фактом в ВЛ-системах, именно, ССС можно классифицировать следующим образом. 1) ССС хранится в словнике системы; 2) ССС не хранится в словнике системы, но сводится к хранимой в словнике ССС алгоритмами морфологического анализа. В настоящее время имеется несколько таких систем; 3) ССС не хранится в словнике системы, но сводится к хранимой в словнике ССС алгоритмами словообразовательного анализа. Системы такого рода находятся в стадии разработки (см., например, 4, 5); 4) ССС не хранится



в словнике системы и не сводится к элементам словника морфологическими и словообразовательными алгоритмами.

Определение 2. *Слово* в ВЛ-системе — это ССС, указанная в п. 1—3 последнего перечня. ССС, указанная в п. 4, называется *псевдословом* в ВЛ-системе.

В ВЛ-системах предусматривается останов, когда встречается псевдослово. Если псевдослово — результат ошибки, то она исправляется. Если ССС — слово в интуитивном смысле, ранее не включенное в словник (например, ПВГДРК тоже может быть словом — это имя персонажа в рассказе Лема, и оно, в принципе, может быть аббревиатурой), то его включают в словарь системы, и оно становится словом в ВЛ-системе.

Алгоритм выделения слова из предложения приведен ниже (рисунок).

1. $k = 1$, def $= 0$:
2. $w = $ —, $i = 1$
3. $t_k \in A$
4. $w_i = t_k$, $i = i + 1$, $k = k + 1$
5. $t_k =$ дефис
6. def $=$ def $+ 1$
7. $t_k = \sigma$
8. $t_{k+1} = \sigma$
9. $k = k + 1$
10. $t_k =$ end t
11. Stop
12. Stop 1
13. $w \Rightarrow$ словник
14. обработка w
15. $k = k + 1$
16. def > 0
17. $j = i$
18. $j = j - 1$, $k = k - 1$
19. $w_j \in A$
20. $w_j =$ дефис
21. $w_j =$ —
22. $k = k + 1$, def $= 0$
23. $t_{k-2} = T8\ t_{k-1} = \cdot 8 t_k = $ —$8\ t_{k+1} = $ $= Д \vee П \vee l8\ t_{k+2} = .$
24. $w_2 =$ точка, $w_3 =$ —, $w_4 = t_{k+1}$, $w_5 =$ точка, $k = k + 3$
25. Stop 2

Обозначения. $A = \{А|а|Б|б|...|Я|я\}$ (русский алфавит), ЗП $= \{.|,|...|„|"|«|»|!|?|:|;|(|—|)|/|\}$ (знаки препинания без дефиса), $t = t_1 t_2 t_3 ...$end$t$ (end$t$ — метка конца текста), $\sigma =$ ЗПU—, def — число дефисов в тексте, def $\geqslant 0$, w — n-местный массив, где помещается слово.

Объяснение блоков. 1°. $k = 1$ означает, что обработка начинается с первой буквы текста, def $= 0$ означает, что пока не встречено дефиса.

2°. Массив w размера n очищается пробелами. В качестве w можно выбрать длину максимального слова в словнике. 3—4°. Если попадаются буквы, то они заносятся в w. 5—6°. Если попадается дефис, то он заносится в w, а в 6° ведется счет дефисов. 7—9°. Слово окончилось; просматривается любое нагромождение символов, пока в t_{k+1} не появится первая буква. 10—12°. При достижении конца файла, содержащего текст в 11°, происходит нормальный останов, останов Stop 1 происходит при наличии непредусмотренного символа в тексте. 13—15°. Блок

13° — сложный. Здесь проверяется, имеется ли w в словнике или сводится ли w к элементам словника с помощью словообразовательных и словоизменительных алгоритмов. Если да, то w обрабатывается в 14°, и в 15° переход на дальнейший просмотр текста. 16°. Проверка, есть ли в w дефис. 17—22°. Массив w проходится от конца к началу, причем дефисы заменяются пробелами. 23—24°. Обрабатываются исключения — слова, содержащие пробел. Это аббревиатуры *т. е., т. д., т. п.* 25°. Встретилось псевдослово. Печать запроса и останов.

Список литературы: 1. *Шенк Р.* Обработка концептуальной информации. —М.: Энергия, 1980. — 361 с. 2. *Звегинцев В. А.* Предложение и его отношение к языку и речи. — М.: Изд-во Моск. ун-та. — 307 с. 3. *Пиотровский Р. Г.* Текст, машина, человек. — Л.: Наука, 1975. — 326 с. 4. *Бондаренко М. Ф., Шаронова Н. В.* Моделирование фрагментированных суффиксов имен существительных. — Депон. рукопись, ВИНИТИ, 1981, № 964. 5. *Бондаренко М. Ф., Лазаренко О. В.* Математическое описание фонетических явлений приставочного словообразования. — АСУ и приборы автоматики, 1981, вып. 58, с. 97—100.

Поступила в редколлегию 25.11.82.