УДК 510.62

А. Ф. ОСЫКА, канд. техн. наук, *И. В. ЗАМАРУЕВА, И. Н. ВОРОНИНА*

НЕКОТОРЫЕ АСПЕКТЫ МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ СЕМАНТИКИ РУССКИХ СЛОВОСОЧЕТАНИЙ

Автоматизация анализа и синтеза текстов на естественном языке требует построения формальных моделей семантики языка [1—3]. Необходимым этапом построения таких моделей является выбор исходных семантических элементов и отношений, в которые вступают между собой эти элементы. Существуют различные подходы к решению задачи нахождения семантических элементов при анализе связных текстов: выделяются единицы значения, соответствующие некоторым объектам реального мира [2]:

— априорно вводятся элементарные единицы значения связанных текстов [1];

— в качестве единицы значения берется лексическое значение слова [4].

Ни один из таких подходов не обеспечивает адекватного моделирования семантики естественного языка. Одна из причин этого состоит в том, что объектом анализа при выделении семантических элементов и отношений являются либо связные тексты [1, 2], либо предложения [4]. Объект слишком сложный, в котором переплетаются явления самых различных уровней языковой системы. Поэтому при его анализе не обойтись без заранее выбранных схем логического, математического или иного характера. В связи с этим представляется целесообразным в качестве исходного объекта анализа с целью выявления элементарных единиц семантического языка и отношений между такими единицами

использовать словосочетания минимальной длины, характерные тем, что они однозначно определяют значение входящих в него слов и описывают некоторый фрагмент реальной действительности. Такой подход используется в некоторых работах для изучения значения отдельных слов [5, 6].

Множество подобных словосочетаний следует разбить на подмножества в соответствии с некоторыми признаками. Внутри каждого подмножества задача выявления семантических элементов и отношений между ними, которые используются в естественном языке, представляется более обозримой, чем в случае анализа связных текстов или предложений. При объединении семантических элементов и отношений, полученных для каждого подмножества, можно получить полный перечень исходных элементов и отношений, реализуемых в текстах интересующего подъязыка.

Рассмотрим случай двухсловных сочетаний. Его состав минимален, поэтому оптимален для анализа, если только значения обоих слов в нем определены полностью. Например, такая ситуация имеет место в словосочетаниях «зеленый мяч», «варить суп», «ножка стула». В словосочетаниях «выпить стакан», «мальчик видит» значения входящих слов и (или) фрагмент реальной действительности, обозначаемый соответствующим словосочетанием, не ясны, причем эта неясность не сводится к незаполненности лексических валентностей слов. В сочетании «варить суп» ситуация ясна, хотя субъект действия неопределен: «некто варит суп». Эта лексическая и, возможно, семантическая валентность заполняется стандартным образом. В случае «вылить стакан» не ясно лексическое значение слова «вылить», и в какой роли используется слово «стакан» — как предмет — объект литья или мера количества жидкости. Для анализа следует отбирать словосочетания первого типа и не отбирать — второго. Но зачастую трудно с уверенностью установить, полно или неполно определены значения слов в словосочетании, так как для этого нет объективных критериев. Вопрос о возможном критерии будет рассмотрен ниже.

В двухсловных словосочетаниях AB значение первого слова A предполагает наличие некоторых семантических черт у слова B, и наоборот. Проследить взаимное влияние значений слов A и B в составе словосочетания удобно, используя матрицу сочетаемости $C=\|C_{ij}\|$ размера $m\times n$. Элементы C_{ij} получают с помощью информанта. Рассмотрим пример такой матрицы на материале словосочетаний AB типа «глагол+существительное в винительном падеже». В качестве значений словоформы A выбраны значения следующих глаголов: 1) готовить, 2) варить, 3) кипятить, 4) жарить, 5) тушить, 6) печь, 7) выпекать, 8) фаршировать, 9) греть, 10) есть, 11) пить. Словоформа B принимает значения таких существительных: 1) еда, 2) борщ, 3) суп, 4) рассольник, 5) компот, 6) чай, 7) кофе, 8) каша, 9) толокно, 10) картофель, 11) морковь, 12) свекла, 13) перец, 14) овощи, 15) мясо, 16) пироги, 17) торт, 18) печенье, 19) хворост, 20) утка, 21) гусь, 22) теленок, 23) животное, 24) птица, 25) обед, 26)завтрак, 27) ужин, 28) тра-

пеза, 29) застолье, 30) угощенье, 31) закуска, 32) десерт, 33) коктейль, 34) питье, 35) бульон, 36) напиток.

Сочетаемость перечисленных глаголов и существительных в значениях, связанных с пищей, может быть передана посредством такой матрицы, строки которой соответствуют существительным, а столбцы — глаголам. Элементы C_{ij} матрицы принимают три значения. Элемент равен 1, если соответствующие глагол и существительное в указанных значениях образуют осмысленное словосочетание (готовить еду, варить картофель и т. д.). $C_{ij}=0$, если глагол и существительное не образуют осмысленное словосочетание (кипятить торт, печь бульон и т. д.). Элемент равен 2, если информант не уверен, можно ли употребить вместе эти два слова с точки зрения стиля, нормы. Но смысл словосочетания вполне понятен (варить десерт, жарить печенье и т. п.).

При образовании всего словосочетания значение X_j некоторого глагола А и значение Y_i конкретного существительного В вступают между собой в определенное отношение, задаваемое предикатом $R_p(X_j, Y_i)$. Способ представления значений X_j и Y_i, реализация механизма взаимодействия между X_j и Y_i по типу R_p на данном этапе исследования не являются существенными. Предикаты $R_p(X_j, Y_i)$ принимают свои значения из множества {0, 1, 2}. В приведенной матрице смежностей указаны конкретные значения предикатов $R_p(X_j, Y_i)=C_{ij}$ ($p=1, 2, 3, \ldots, j=1, 2, \ldots, 9, i=1, 2, \ldots, 36$) для каждой пары значений X_j и Y_i.

Нет оснований изначально считать, что из приведенных выше 9 глаголов и 36 существительных в заданных значениях всегда реализуется одно и то же отношение между значениями X_j и Y_i. Поэтому введено множество отношений $R_p(X_j, Y_i)$ ($p=1, 2, 3\ldots$). Например, с точки зрения носителя языка одинаковы отношения между X_2=«варить» и Y_{10}=«картофель» в словосочетании «варить картофель» и X_5=«тушить (пищу)» и Y_{11}=«морковь» в словосочетании «тушить морковь», т. е. $R_p(X_2, Y_{10})=1$ и $R_p(X_5, X_{11})=1$.

Для других пар X_j и Y_i эти отношения воспринимаются как различные: X_8=«фаршировать» и Y_{13}=«перец» в словосочетании «фаршировать перец» в отличие от X_{11}=«пить», Y_6=«чай» в — «пить чай». Здесь $R_q(X_8, Y_{13})=1$ и $R_k(X_{11}, X_6)=1$, $q\neq k$.

Важен вопрос о том, какие сведения о значениях X_j и Y_i и связях между ними можно получить из факта сочетаемости этих значений, т. е. наличия этих значений у синтаксически связанных слов (или просто рядом расположенных). Иными словами, важно выявить, в какой мере интуитивные представления о значениях X_j и Y_i и отношениях между ними $R_p(X_i, Y_j)$ проявляются через сочетаемость этих значений в составе словосочетаний минимальной длины. Рассмотрим этот вопрос на материале приведенной матрицы сочетаемости. Естественно предположить, что если несколько существительных сочетается с некоторым глаголом со значением X_j, то специфика значений этих существительных обуславливает наличие определенных семантических элементов в значении X_j. Например, X_2=«варить» сочетается с «суп», «рассоль-

	1	2	3	4	5	6	7	8	9	10	11
1	1	1	0	1	1	2	2	2	1	1	0
2	1	1	1	0	0	0	0	0	1	1	0
3	1	1	1	0	0	0	0	0	1	1	0
4	1	1	1	0	0	0	0	0	1	1	0
5	1	1	1	0	0	0	0	0	1	0	1
6	1	1	1	0	0	0	0	0	1	0	1
7	1	1	1	0	0	0	0	0	1	0	1
8	1	1	0	0	0	0	0	0	1	1	0
9	1	1	0	0	0	0	0	0	1	1	0
10	1	1	0	1	1	1	0	1	1	1	0
11	1	1	0	1	1	1	0	1	1	1	0
12	1	1	0	1	1	1	0	1	1	1	0
13	1	1	0	1	1	2	0	1	1	1	0
14	1	1	0	1	1	2	0	2	1	1	0
15	1	1	0	1	1	1	0	1	1	1	0
16	1	0	0	2	0	1	1	2	1	1	0
17	1	0	0	0	0	1	1	0	2	1	0
18	1	0	0	2	0	1	1	0	2	1	0
19	1	0	0	2	0	1	1	0	2	1	0
20	1	1	0	1	1	1	0	1	1	1	0
21	1	1	0	1	1	1	0	1	1	1	0
22	1	1	0	1	1	1	0	1	1	1	0
23	1	1	0	2	2	1	0	1	1	1	0
24	1	1	0	1	1	1	0	1	1	1	0
25	1	1	0	1	0	0	0	0	1	1	0

	1	2	3	4	5	6	7	8	9	10	11
26	1	1	0	1	0	0	0	0	1	1	0
27	1	1	0	1	0	0	0	0	1	1	0
28	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0
30	1	1	0	1	2	2	2	0	1	1	1
31	1	2	0	2	2	2	2	0	1	1	0
32	1	2	0	2	2	2	2	0	1	1	0
33	1	2	2	0	0	0	0	0	1	0	1
34	1	1	1	0	0	0	0	0	1	0	1
35	1	1	1	0	0	0	0	0	1	2	1
36	1	1	1	0	0	0	0	0	1	0	1

ник», «чай», «кофе», «картофель» и др. При этом в значении X_2 выделяются такие элементы: «изготовление пищи», «изготовление путем нагревания», «кипячение жидкости» и др. Существительное со значением Y_{10}=«картофель» сочетается, в частности, с глаголами «варить», «жарить», «тушить», «печь», которые выделяют в значении Y_{10} такие элементы: «пищевой полуфабрикат», «изготовление путем нагревания» и т. п. Таким образом, с некоторой долей условности можно считать, что если $C=\|C_{ij}\|$ — матрица сочетаемости размера $m\times n$, то ее j-й столбец $(C_{1j}, C_{2j}, \ldots, C_{mj})$ — это характеристика значения X_j соответствующего глагола, а i-я строка $(C_i, C_{i2}, \ldots, C_{in})$ — характеристика значения Y_i соответствующего существительного.

Чем больше элементов C_{ij} в столбце j равно 1 или 2 для глагола со значением X_j, тем более широким лексическим значением, как правило, обладает данный глагол, тем меньше специфических черт — элементарных единиц значения входит в значение X_j, чтобы не ограничивать сочетаемость с разнообразными существительными. В приведенной матрице наибольшее количество единиц и двоек имеется в 1, 2, 9 и 10 столбцах, которые соответствуют значениям глаголов «готовить», «варить», «греть», «есть».

Аналогичное замечание можно сделать относительно строк матрицы. Наибольшее количество единиц и двоек имеется в строках 1, 12, 14, 21, 24, которые соответствуют значениям существительных «еда», «свекла», «овощи», «гусь». Если имеются значения двух глаголов X_j и X_q, причем X_j сочетается только с частью зна-

чений существительных, с которыми сочетается X_q, и только с ними, то можно предположить, что X_q является более общим значением в сравнении с X_j. Иными словами, X_q можно рассматривать как родовое значение по отношению к видовому $X_j (X_j \Rightarrow X_q)$. В терминах элементов матрицы сочетаемости это предположение формулируется так:

$$\exists (C_{1j}, C_{2j}, \ldots C_{nj}), \exists (C_{1q}, C_{2q}, \ldots C_{nq}), \forall_i (C_{ij} = \overline{0} \supset C_{iq} = 1) \supset \supset X_j \Rightarrow X_q.$$

Аналогичное замечание правомерно и относительно связи между строками матрицы сочетаемости и значениями Y_i и Y_k существительных:

$$\exists (C_{i1}, C_{i2}, \ldots C_{im}), \exists (C_{k1}, C_{k2}, \ldots C_{km}), \forall j (C_{ij} = \overline{0} \supset C_{kj} = 1) \supset \supset Y_i \Rightarrow Y_k.$$

Проверим, насколько значение элементов матрицы сочетаемости согласуется с этими предположениями. Единичные элементы столбца 1 («готовить») полностью включают в себя ненулевые элементы столбцов 2, 4, 5, 6, 7, 8 («варить», «жарить», «тушить», «печь», «выпекать», «фаршировать»). Это вполне согласуется с представлениями информанта о соотношении значений как видовое и соответствующее родовое: X_2 = «варить» $\Rightarrow X_1$ = «готовить», ..., X_8 = «фаршировать» $\Rightarrow X_1$ = «готовить».

Имеются и другие случаи вложения всех ненулевых элементов одних столбцов в единичные элементы других столбцов, которые сопровождаются родо-видовыми соотношениями между значениями глаголов. Например, столбец 5 («тушить») полностью вкладывается в столбец 4 («жарить») и 2 («варить»), а столбец 7 («выпекать») — в 6 («печь»). Важно отметить, что нет случаев, когда родо-видовые соотношения между значениями глаголов не сопровождаются вложением соответствующих столбцов матрицы сочетаемости.

Вместе с тем имеются случаи вложения столбцов, которым не соответствует родо-видовая связь между значениями глаголов. Например, столбец 1 («готовить») включает в себя столбцы 9 («греть»), 3 («кипятить») и 11 («пить»), а столбец 8 («фаршировать») входит в столбцы 1 («готовить»), 2 («варить»), 4 («жарить»), 5 («тушить»), 6 («печь»), 9 («греть»), 10 («есть»). Эти случаи рассогласования сочетаемости и значений глаголов можно разбить на два типа в зависимости от того, какими средствами это несоответствие устраняется. Ложное вложение столбцов первого типа возникает из-за того, что учтены не все существительные, специфицирующие значения глаголов. Например, включение столбца 9 в столбец 1 устраняется добавлением к списку существительных слова «камень». Тогда получаем «готовить камень» — не сочетается, а «греть камень» — сочетается. При этом элемент значения «приготовление пищи» выводится из значения глагола «греть». Вложение 3-го столбца в 1-й устраняется добавлением существительного «кислота», а в 1-й — добавлением «вода»,

Неадекватное вложение столбца второго типа не может быть устранено путем пополнения списка существительных. Например, любой продукт, который можно фаршировать, можно вместе с тем и готовить, и варить, и жарить, и тушить и т. п. Необходимо расширение контекста для уточнения значения глагола. Такие случаи вполне естественны, так как исходный материал для семантического анализа отбирался по синтаксическому критерию. Вовсе не обязательно, чтобы выбранная синтаксическая конструкция минимальной длины содержала все слова, существенные для спецификации значения данного глагола.

Устранить ложное вложение столбца 8 в столбцы 1, 2, 4, 5, 6, 9 и 10 можно при рассмотрении расширенного словосочетания типа «фаршировать мясо овощами». С таким распространением ни один из глаголов 1, 2, 4, 5, 6, 9 и 10 не сочетается. Глаголы, подобные «фаршировать», необходимо анализировать в составе сочетаний, содержащих три или более слов. Для описания семантики таких словосочетаний следует использовать трех (или более)-местные элементарные предикаты.

Все наблюдения, касающиеся связи значений глаголов и соответствующих столбцов в матрице сочетаемости, в равной мере справедливы и для значений существительных, и связанных с ними строк матрицы. Дальнейшие сведения о соотношениях между значениями слов могут быть получены в результате анализа случаев, когда ненулевые элементы двух столбцов (строк) совпадают частично или не совпадают вовсе.

Таким образом, необходимым этапом формального описания семантики слов по их взаимной сочетаемости и семантики словосочетаний является правильный отбор исходного материала для анализа. Критерий отбора словосочетаний — соответствие строк и столбцов в матрице сочетаемости интуитивным представлениям носителя языка о соотношении между значениями слов.

Если в матрице имеется вложение столбцов (строк) и нет родо-видовых отношений между значениями соответствующих слов, то необходимо либо расширить список сочетающихся слов для устранения ложного вложения, либо описание семантики словосочетания двухместным предикатом невозможно.

Между некоторыми значениями глаголов, существительных устанавливаются родо-видовые отношения. В таком случае они могут рассматриваться в качестве семантической переменной, используемой в записи предиката словосочетания, а соответствующие видовые значения — в качестве значений переменной.

Полученные выводы проверены на материале некоторых других совокупностей словосочетаний.

Список литературы: 1. *Шенк Р.* Обработка концептуальной информации. М., 1980. 358 с. 2. *Новиков А. И.* Семантика текста и ее формализация. М., 1983, 214 с. 3. *Кузнецов И. П.* Механизм обработки семантической информации. М., 1978. 174 с. 4. *Арутюнова Н. Д.* Предложение и его смысл. М., 1976. 356 с. 5. *Апресян Ю. Д.* Экспериментальное исследование семантики русского глагола. М., 1976. 262 с. 6. *Апресян Ю. Д.* Лексическая семантика. М., 1974. 324 с.

Поступила в редколлегию 30.03.87