УДК 519.766.2

Ю. Ю. ЧЕРЕПАНОВА

О ТЕОРЕТИКО-МНОЖЕСТВЕННОМ И ТЕОРЕТИКО-КАТЕГОРНОМ ПОДХОДАХ К МОДЕЛИРОВАНИЮ СЕМАНТИЧЕСКИХ ПОЛЕЙ

В лингвистике широкое распространение получил полевой подход, в рамках которого семантическое пространство естественного языка разбивается на множество семантических полей [1-5]. В проведенном анализе семантических полей [6, 7] было установлено, что их структура включает в себя лексические единицы различных уровней и связывающие их семантические отношения. Такое строение семантических полей дает возможность моделирования их с помощью математического аппарата теории множеств, теории графов и теории категорий.

Теоретико-множественное моделирование.

Формализуем лингвистические модели семантических полей с помощью теоретико-множественного аппарата.

Определим следующие множества:

$V = \{v_1, v_2, \ldots, v_M\}$ – множество слов языка;

$A = \{a_1, a_2, \ldots, a_K\}$ – множество семантических признаков;

$T = \{t_1, t_2, \ldots, t_N\}$ – множество семантических типов.

Множества V, A, T представляют собой лексические единицы различных уровней. Для отражения существующих между ними семантических связей введем следующие отношения.

Определим отображение $f_{код}: V \to A$, где $V = \{v_1, v_2, \ldots, v_M\}$ – множество слов языка, $A = \{a_1, a_2, \ldots, a_K\}$ – множество семантических признаков. Данное отображение представляет собой кодирование слов в квазиосновы, семантические признаки. Это отображение является всюду определенным, сюръективным. Так как один признак может соответствовать нескольким словам (отображение не является инъективным), $K \le M$.

Отношение $R_{см} \subseteq A \times A$ содержательно означает «иметь семантический множитель». Это отношение всюду определенное (так как у любого слова есть определение, следовательно можно построить семантический объем), рефлексивное (так как в семантический объем включается само слово), транзитивное (семантические множители включаются в объем итерационно). Тогда сечения $R_{см}(a_i)$ представляют собой семантические объемы слов $v_i = f_{код}^{-1}(a_i)$.

При построении семантических полей все множество слов V разбивается на множество семантических полей $V_1, V_2, \ldots, V_N$. Семантическое поле V_i – некоторое множество слов $\{v_1, v_2, \ldots, v_M\}$, связанных между собой парадигматическими и синтагматическими семантическими отношениями; соответственно разделяются парадигматические и синтагматические семантические поля. К парадигматическим полям относятся различные классы лексических единиц, имеющие общие признаки в своих значениях. Следовательно, если каждому слову v_i поставить в соответствие множество $W_i = R_{см}(a_i) = \{a_{i1}, a_{i2}, \ldots, a_{iK}\}$ (семантический объем), где a_{iL} – семантический признак слова v_i, то слова v_i, v_j объединяются в одно парадигматическое поле U_z при условии, что $W_i \bigcap W_j \neq \varnothing$.

Таким образом, можно задать отношение $R_n \subseteq V \times V$, определяющее наличие парадигматических связей между словами. Отношение R_n всюду определенное, сюръективное, ре-

флексивное, симметричное. Сечения $R_n(v_i)$ представляют собой парадигматические поля слов v_i.

Дадим формальное определение некоторых видов парадигматических полей и отношений. Пусть слову v_i соответствует множество семантических признаков $W_i = \{a_{1i}, a_{2i}, \ldots, a_{Mi}\}$, слову v_j множество $W_j = \{a_{1j}, a_{2j}, \ldots, a_{Kj}\}$. Тогда для определения вхождения слов v_i, v_j в поле U необходимо проанализировать следующее множество: $W = W_i \cap W_j = \{a_1, a_2, \ldots, a_Z\}$.

Лексико-семантическая группа определяется как группа слов, в семантических объемах которых совпадает хотя бы один семантический множитель, тогда: $W \neq \varnothing \Rightarrow v_i, v_j \in U$.

При построении тематического ряда фиксируются определенные семантические признаки (путем выбора одного конкретного слова и его семантического объема или отдельных семантических множителей). Значит: $W_T = \{a_{T1}, a_{T2}, \ldots, a_{TL}\}, L \geq 1, W_T \subseteq W_i \Rightarrow v_i \in U$.

В синонимические ряды включаются слова, семантические объемы которых совпадают более чем на 50%. Следовательно, формальное условие вхождения слов v_i, v_j в один синонимический ряд выглядит так: $Z \geq \frac{1}{2} min(M,K) \Rightarrow v_i, v_j \in U$.

При абсолютных синонимах семантические объемы должны быть одинаковы: $W_i = W_j \Rightarrow v_i, v_j \in U$.

Необходимым (но не достаточным) условием того, что v_i и v_j находятся в родовидовых отношениях (v_i является видом v_j), является: $W_i \subseteq W_j$. Это объясняется тем, что значение родового понятия (более центрального в поле), не может быть уже значения видового понятия, следовательно, его семантический объем будет меньше и включается в семантический объем видового понятия.

Синтагматическими называются отношения между сочетаемыми единицами языка. Таким образом, синтагматические семантические поля слова представляют собой связи слов либо их абстрактных категорий в текстах. Для отражения этих связей необходимо ввести несколько отношений.

Введем всюду определенное, сюръективное отношение $R_{cm} \subseteq V \times T$, означающее «сочетаться с семантическим типом».

Отношение $R_{mm} \subseteq T \times T$, определяющее сочетание семантических типов, всюду определено, сюръективно, симметрично, антирефлексивно.

Введем отображение $f_m : V \to T$, содержательно означающее «относиться к семантическому типу».

Тогда отношение $R_{cn} \subseteq R_{cm} f_T^{-1}$ определяет синтагматические связи между словами. Сечения $R_{mm}(t_j)$ представляют собой синтагматические связи категории t_j, сечения $R_{cm}(v_i)$ – синтагматические связи слов v_i с абстрактными категориями в текстах, а сечения $R_{cn}(v_i)$– связи слов v_i с другими словами.

Разумеется, данная модель является очень упрощенной. Отношения R_n и R_{cn} отражают лишь наличие соответствующих связей между словами. На самом деле парадигматические и синтагматические связи должны отражаться множеством отношений. Кроме того, данная модель отражает многозначность лексических единиц лишь в том случае, если разные значения слов отображаются как разные слова v_i, v_j.

Теоретико-категорное моделирование. По определению [8, 9], категория K состоит из двух непересекающихся классов элементов: класса объектов $Ob\,K$ и класса морфизмов $Mor\,K$, которые удовлетворяют следующим аксиомам:

1. Задано отображение $H_K : Ob\,K \times Ob\,K \to P(Mor\,K)$, которое разным парам объектов (A, B) и (A', B') сопоставляет непересекающиеся множества морфизмов $H_K(A, B)$ и $H_K(A', B')$.

2. В каждом множестве $H_K(A, A)$ выделяется специальный морфизм 1_A.

3. Для любых двух морфизмов $f \in H_K(A, B),\ g \in H_K(B, C)$ определено их произведение $fg \in H_K(A, C)$.

4. Для любых трех морфизмов $f \in H_K(A, B),\ g \in H_K(B, C),\ h \in H_K(C, D)$ выполняется равенство $(fg)h = f(gh)$.

5. Для любого морфизма $f \in H_K(A, B)$ выполнены равенства $1_A f = f = f1_A$.

В таком случае, моделирование семантических полей с использованием математического аппарата теории категорий возможно, если при построении категорий семантических полей в качестве объектов выступают лексические единицы, в качестве морфизмов - семантические отношения между лексическими единицами. Специальным морфизмом, определенным в аксиоме 2, является отношение тождественности лексической единицы самой себе, а правилам композиции морфизмов, определяемым в аксиомах 3 и 4, соответствуют правила получения семантических отношений.

Приведем некоторые примеры теоретико-категорного моделирования семантических полей.

Категория Π_{vi}. Определим категорию, моделирующую семантический объем лексической единицы v_i.

Согласно приведенному формальному определению семантического объема слова, объектами категории являются семантические признаки, морфизмами – отношения «являться семантическим множителем». При этом объекты данной категории могут быть определены индуктивно:

1. Семантический признак $a_l = f_{код}(v_i)$ является объектом категории Π_{vi}.

2. Если $a_y \in Ob\,\Pi_{vi}$, и a_j является семантическим множителем a_y, то $a_j \in Ob\,\Pi_{vi}$.

Количество шагов индукции определяется требуемой степенью детализации компонентного анализа.

Так как выполняется аксиома 1 (семантический признак входит в свой семантический объем на следующем шаге компонентного анализа), определено произведение морфизмов (семантический множитель семантического множителя также является семантическим множителем), и произведение морфизмов ассоциативно, $\Pi_{vi} = \langle Ob\,\Pi_{vi}, Mor\,\Pi_{vi} \rangle$ действительно представляет собой категорию. Это связанная тонкая категория, имеющая хотя бы один инициальный объект $a_l = f_{код}(v_i)$. Если в категории Π_{vi} кроме объекта $a_l = f_{код}(v_i)$, существуют и другие инициальные объекты a_j, то они изоморфны, и слова v_i и $v_j = f_{код}^{-1}(a_j)$, являются синонимами.

При построении парадигматического поля для определения наличия парадигматического семантического отношения между словами v_i, v_j необходимо сравнить категории Π_{vi} и Π_{vj}, а именно $Ob\,\Pi_{vi} \bigcap Ob\,\Pi_{vj}$, т. е. найти соответствие между $Ob\,\Pi_{vi}$ и $Ob\,\Pi_{vj}$.

Категория Π_M. Определим теоретико-категорную парадигматическую модель некоторого текста (аналогичную упрощенной концептуальной модели текста или предметной области, описанной в [7]). Такой моделью является объединение категорий Π_{vi} значащих слов текста или выбранных понятий, а именно категория $\Pi_M = \bigcup_{i=1}^{n} \Pi_{vi}$, такая, что

$Ob\ П_м = \bigcup_{i=1}^{n} Ob\ П_{vi}$, и $Mor\ П_м(a,b) = \bigcup_{i=1}^{n} Mor\ П_{vi}(a,b)$ для каждой пары $a,b \in Ob\ П_м$.

Таким образом, каждая категория $П_{vi}$ является подкатегорией $П_м$.

Расширение категории $П_м$ по всей предметной области (по всем значащим лексическим единицам) даст модель парадигматического тезауруса.

Категория $П_{мр}$. Расширенная концептуальная модель некоторого текста [7] представляет собой категорию $П_{мр}$, объектами которой являются категории $П_{vi}$, где v_i – слова в тексте, а морфизмами – реализованные синтагматические отношения в тексте.

Расширение данной модели по всей предметной области (построение по достаточно широкому классу текстов данной предметной области) представляет собой категорную модель тезауруса предметной области. В этой категории $П_{по}$ объектами являются категории $П_{vi}$, где v_i – все значащие слова предметной области, а морфизмами – все синтагматические отношения в тексте.

Для установления адекватности $П_{мр}$ модели $П_{по}$ необходимо установление соответствий между $Ob\ П_{мр}$ и $Ob\ П_{по}$, $Mor\ П_{мр}$ и $Mor\ П_{по}$. Если $П_{мр} \subseteq П_{по}$,то соответствие полное.

Список литературы: 1. *Щур Г. С.* Теория поля в лингвистике. - М.: «Наука», 1974. - 255 с. 2. *Васильев Л. М.* Современная лингвистическая семантика. – М.: Высш. шк., 1990. – 176с. 3. *Кузнецов А. М.* Структурно-семантические параметры в лексике. – М.: Наука, 1980. - 160с. 4. *Караулов Ю. Н.* Лингвистическое конструирование и тезаурус литературного языка. – М.: Наука, 1981. – 367с. 5. Полевые структуры в системе языка /[З. Д. Попова, И. А. Стернин, Е. И. Беляева и др.] Науч.ред. З. Д. Попова. – Воронеж: Изд-во ВГУ, 1989. – 196с. 6. *Павлов П. Ф., Бабина О. И., Черепанова Ю. Ю.* О проблеме автоматизированного выявления семантических полей. // Вісник ХДПУ. Збірка наукових праць. – Вип. 42. – Харьков: ХДПУ, 1999. – С. 81-85. 7. *Павлов П. Ф., Черепанова Ю. Ю. Шубин И. Ю.* О возможности построения тезауруса семантических полей и его применения в информационных системах // Вісник ХДПУ. Збірка наукових праць. – Вип. 108. – Харьков: ХДПУ, 2000. – С. 41-46. 8. *Цаленко М. Ш.* Моделирование семантики в базах данных. – М.: Наука. Гл. ред. физ-мат. лит., 1989. – 288с. 9. Общая алгебра. Т 2./ Под ред. Л. А. Скорнякова.– М.: Наука. Гл. ред. физ-мат. лит., 1991. – 480с.

Поступила в редакцию
28.06.2001