Д.Н. Гомон (Минск, БГУ)

ВЕРОЯТНОСТНАЯ МОДЕЛЬ ОМОНИМИИ

Факты того или иного языка можно изучать в нескольких планах: лексемы, зафиксированные в *словаре* (**С**);– лексемы и словоформы, обнаруженные в текстах — *текстуальные* (**Т**); словоформы, которые пока или уже не находятся в доступных текстах и словарях, но разрешены системой языка и могут быть нами обнаружены либо смоделированы, — *гипотетические* (**Г**). Универсумом для данной классификации может выступать как лексика языка в общем, так и омонимика в частности.

Подсистема 8 (– — –) содержит универсум до начала его классификации.

Подсистема 5 (+ — –) содержит омонимы, обнаруженные только в словарях, но не встреченные в текстах. Для латинского языка это, к примеру, словарь омонимов в словарной форме, составленный нами [1, с. 135–193].

Подсистема 6 (– + –) содержит множество текстов данного языка.

Подсистема 7 (– — +) будет содержать словоформы, которых нет в доступных текстах и словарях, но они разрешены системой языка и могут быть обнаружены либо созданы нами. Слово *запил* как результат действия по глаголу *запилить* не обнаружено ни в одном современном словаре русского языка.

Однако его существование возможно, т. к. нет запрета на совместное употребление приставки *за-* и основы *пилить*. На это указывают и другие однокоренные омонимы (таблица 1.), отличающиеся приставкой:

Таблица 1

Гомологический ряд с гоморазницей в приставку

пить	запить	отпить		испить	пропить	Распить	
пилить	запилить	отпилить		испилить	пропилить	распилить	
пил	*запил	отпил*			пропил*	распил*	
пропил					пропил		
пой	запой*		напой* (бел)		пропой*		припой
поить		отпоить	напоить				
петь	запеть	отпеть*	напеть*		пропеть		

В табл. 1 мы включили только словарные формы слов, которые образуют лексико-грамматическую (в некоторых случаях и лексическую) омонимию. Данная таблица представляет собой вариант гомологического ряда, в котором гомологическую разницу составляют приставки (в столбцах). В строках имеем только однокоренные слова. Видим, что наличие в системе существительных *пропил* (от *пропилить*) и *распил* создает технологию для создания по системному образцу иных существительных от глагола *пилить*. В случае обнаружения слова *запил* в тексте оно перейдет в подсистему 4; а если оно будет найдено в словаре — в подсистему 3 или 1. К слову, данная лексема была обнаружена нами в словаре Даля [2, с. 1534].

Подсистема 4 (– + +) включает тексты, составленные с применением разрешенных правилами языка гипотетических словоформ. Практически это множество текстов на жаргоне и арго, сленге, результат речемыслительной деятельности осваивающих язык (детей, иностранцев).

Подсистема 3 (+ — +) включает признаки словарных и гипотетических фактов. Считаем, что это словник всех разрешенных лексем определенного языка (словарь возможных миров, содержащий, к примеру, слова *взничь*, *припрыжка*, *льзя* и другие им подобные; словарь ошибочно составленных детьми или иностранцами слов; словарь «народной энтимологии»).

Подсистема 2 (+ + –) содержит всю наличную систему омонимики языка без учета «семантики возможных миров».

Подсистема 1 (+ + +) содержит омонимику языка, а также разрешенные правилами языка возможные словоформы — язык-систему.

Понятие лингвистическая «семантика возможных миров» [4, с. 192–196] предполагает полную теоретическую комбинаторику значимых единиц языка и в этом виде представляет собой базу для предсказывания омонимии.

При теоретическом представлении омонимии важным моментом является вероятностная модель явления, которая, с одной стороны, должна показывать детерминированные мощностью (а значит, разнообразием) частей речи разноплановые (одноуровневые, разноуровневые и т. п.) проявления омонимии, определяемые без использования знания о семантике и грамматике слов, создающих омонимические отношения, а с другой стороны, должна вскрывать неявные запреты (ограничения) на создание указанных отношений. Анализ литературы по проблеме вероятностных моделей в омонимии показывает, что данный подход лингвистами не применялся.

В работе «Язык как система» была построена вероятностная модель для синтаксиса (ядерные конструкции) [4, 102-116]. Количественный изоморфизм (два слова, необходимых для создания омопары, и два слова, необходимых для создания ядерной композиции — словосочетания, предложения) и мощность словаря, на базе которого строилась эта модель (50 тыс. слов), позволяют построить вероятностную модель омонимии латинского языка по аналогии с вероятностной моделью синтаксиса.

При построении нашей вероятностной модели омонимии будем исходить из восьми классов слов, сводя частицы, союзы, междометия в один х-класс в силу их неграмматичности. Исследуются словарные (репрезентативные) формы, отобранные сплошной выборкой из [3]. Теоретически не налагается запрет на то, что два элемента одного или разных классов (частей речи) могут составить омопару (войти в омонимические отношения). Мы предполагаем, что такие ограничения (не запреты) в системе существуют и будут выявлены при сопоставлении вероятностной модели с реальными данными. Для упрощения расчетов данные по частям речи округлены, что дало следующие количественные характеристики:

Subst.	Ns	20000	Praep.	Npraep	80
Verb.	Nv	15000	Num.	Nn	50
Adj.	Na	12000	Pron.	Npr	50
Adv.	Nadv	2700	X-class	Nx	120

Общее число элементов всех классов N = Ns + Nv + Na + Nadv + Npraep + Nn + Npr + Nx = 50 тыс. слов. Вероятности того, что представитель того или иного класса вступит в отношения омонимии (с представителем того же или другого класса), соответственно равны:

Ps = Ns/N = 0.4
Pv = Nv/N = 0.3
Pa = Na/N = 0.24
Padv = Nadv/N = 0.054
Ppraep = Npraep/N = 0.0016
Pn = Nv/N = 0.0010
Ppr = Npr/N = 0.0010
Px = Nx/N = 0.0024

При этом P = Ps + Pv + Pa + Padv + Ppraep + Pn + Ppr + Px = 1.

Вероятность встречаемости представителя данного класса с представителем другого конкретного класса будет представлять произведение вероятностей данных классов. Порядок следования элементов в омопаре не имеет значения, поэтому все вероятности занесены в треугольную декартову матрицу (см. табл. 2).

Таблица 2

Вероятностные данные межклассовой омонимии

	subst.	verb.	adj.	adv.	praep.	num.	pron.	x-класс
subst.	0.16	0.12	0.096	0.0215	0.00064	0.0004	0.0004	0.00096
verb.		0.09	0.072	0.0162	0,00048	0.0003	0,0003	0,00072
adj.			0.0576	0.01296	0,000384	0.00024	0,0024	0,000576
adv.				0.00246	$8{,}64\text{x}10^{-5}$	$5{,}4\text{x}10^{-5}$	$5{,}4\text{x}10^{-5}$	0,00013
praep.					$2{,}56\text{x}10^{-6}$	$1{,}6\text{x}10^{-6}$	$1{,}6\text{x}10^{-6}$	$3{,}84\text{x}10^{-6}$
num.						$1\text{x}10^{-6}$	$1\text{x}10^{-6}$	$2{,}4\text{x}10^{-6}$
pron.							$1\text{x}10^{-6}$	$2{,}4\text{x}10^{-6}$
x-класс								$5{,}76\text{x}10^{-6}$

При умножении вероятности омонимизации каждой пары слов на общее количество слов в словаре получим вероятное количество омопар. Все данные внесем в таблицу 3. В числителе приведем количество вероятных омопар, в знаменателе — реальные данные. Процент приводим к числу словарных статей.

Таблица 3

Количество вероятных и реальных событий

	subst.	verb.	adj.	adv.	praep.	num.	Pron.	x-класс	всего	%
subst.	8000/748	6000/129	4800/1633	1225/75	32/7	20/1	20/3	48/6	2501	12,5
verb.		4500/152	3600/2	810/52	24/0	15/1	15/0	36/1	336	2,2
adj.			2880/195	648/20	19/1	12/0	12/0	29/0	1851	15,4
adv.				123/4	4/33	3/0	3/8	7/12	204	7,3
praep.					0/0	0/0	0/0	0/3	44	55
num.						0/0	0/0	0/0	2	4
pron.							0/1	0/1	12	24
x-кл.								0/1	24	20

В табл. 3 выделяются 3 зоны:

1. Вероятностные данные выше реальных. Таких клеток большинство. Это объяснимо, ведь на 50 тыс. словарных статей и потенциальных омонимов имеем всего 6170 омонимичных словарных статей (около 12 %). При этом более половины слов имеют два и более значения (являются неоднозначными), то есть потенциально представляют собой лексические омонимы. Среди них есть клетки, где реальные данные равны нулю. В основном это числительные, местоимения и представители х-класса, немногочисленность и специфичность которых в значении и, особенно, в плане выражения замедляют омонимизацию.

2. Вероятностные данные ниже реальных. Это наречия с предлогами, местоимениями и представителями х-класса (союзами), которая с очевидностью показывает условную взаимозаменяемость этих частей речи. То, что более 40% предлогов омонимичны наречиям, только доказывает это.

В четырех подсистемах вероятность равна нулю, но реальные примеры имеются (pron.+pron., pron.+х-класс, х-класс+х-класс, praep.+х-класс). Эти клетки показывают противоречие между вероятным и реальным количеством, указывая на значение качественной стороны (хотя бы симметрия одного знака).

3. Вероятностные и реальные данные равны нулю. Ожидавшаяся низкая корреляция (вплоть до ее отсутствия) реализовалась из-за малого числа единиц и (как предполагается) специфичности (неграмматичности или особых способов словоизменения) плана выражения служебных частей речи, междометий, местоимений и числительных.

Между частями речи чаще всего встречаются предлоги (55%), местоимения (24%), представители х-класса (союзы — 20%), прилагательные (15,4%), существительные (12,5%). Реже всего — глаголы (2,2%). Служебные части речи имеют наибольшую омонимичность с наречиями (это обусловлено еще и общностью их происхождения), существительные и прилагательные — между собой. Т.е. преобладает функциональная (конверсионная) омонимия.

Итак, в данной статье мы показали зависимость омонимии от количественных характеристик части речи. Если в расчет вероятности взять другие критерии (частотность употребления слова, величину парадигмы, среднюю длину слова в пределах части речи), то, возможно, результаты окажутся несколько иными. Это может стать новым прикладным направлением в исследовании, равно как и изучение системы противоречий-непротиворечий в процессе возникновения омонимии.

ЛИТЕРАТУРА

1. Гомон, Д.Н. Системные проблемы лексической омонимии: дис. ...канд. филол. наук: 10.02.19. — Минск, 2004.

2. Даль, В. Толковый словарь живого великорусского языка: в 4 т. — М., 1994. — Т.1: А-З.

3. Дворецкий, И.Х. Латинско-русский словарь: 4-е издание, стереотипное — М., 1996.

4. Карпов, В.А. Язык как система. — Минск, 1992.