

Национальный корпус русского языка и проблемы гуманитарного образования.
Материалы международной научной конференции. Москва 19-20 апреля 2007. М.
2007. С. 61-63.

К разработке нового словаря функциональных омонимов на основе Национального корпуса русского языка

О.А. Невзорова
Татарский государственный гуманитарно-педагогический университет,
НИИММ им. Н.Г. Чеботарева
Olga.Nevzorova@ksu.ru

Ю.В. Зинькина
Казанский государственный университет
zjuliv@mail.ru

Одной из наиболее сложных проблем компьютерной лингвистики является проблема разрешения многозначности, которая проявляется в различных аспектах – морфологическом, синтаксическом, лексическом. Одним из методов разрешения функциональной омонимии для русского языка является метод синтаксического анализа ближайшего контекста омонима. В основе метода контекстного анализа лежат эвристические правила, которые учитывают наиболее типичные для русского языка контекстные ситуации для каждого типа функциональных омонимов. Очевидно, что точность метода напрямую зависит от разнообразия разрешающих контекстов, а также типа омонима, который определяется на основе приписанной омониму совокупности грамматических характеристик. Представляется, что только на основе корпусных исследований можно решить поставленные задачи.

При работе с Национальным корпусом русского языка (НКРЯ) решались следующие основные задачи:

1) Получить в ходе корпусных исследований на материале НКРЯ новую классификацию внутри классов грамматических омонимов, в соответствии с особенностями употребления их в современном русском языке. В результате было выявлено три группы омонимов:

- омонимы, употребляющиеся в роли слов разных частей речи - примерно поровну, так называемые регулярные омонимы;

- омонимы, чаще используемые как представители какой-либо одной части речи – нерегулярные омонимы.

- омонимы, встречающиеся в роли представителей только одной части речи - потенциальные омонимы.

Построенная классификация учитывается при разработке правил синтаксического разрешения омонимов. Для каждой подгруппы внутри одного омонимического класса разрабатывается свой набор правил, при этом различаются как отдельные элементы правил, так и их иерархия.

2) Построить максимально точное лексикографическое описание функциональных омонимов. Действительно, на сегодняшний день нет специализированного словаря, охватывающего все функциональные омонимы и приписывающего им однозначные характеристики. Существующие ресурсы часто представляют лишь какую-либо часть омонимов и описывают их порой весьма противоречиво.

Был проведен лексикографический эксперимент, в ходе которого сопоставлялись характеристики, приписываемые 4 ресурсами 560 омонимам на 'о' типа «краткое прилагательное – наречие», с возможной принадлежностью к какому-либо еще морфологическому классу. Лишь в трех случаях все четыре ресурса приписали омонимам одинаковые характеристики. Это были омонимы '*прямо*' и '*исключительно*' с характеристиками «краткое прилагательное – наречие – частица», а также омоним '*относительно*' с характеристиками «краткое прилагательное – наречие – предлог».

При этом следует отметить, что характеристики, предлагаемые НКРЯ, во многих случаях были наиболее точными. Они также в наибольшей степени отвечали требованиям синтаксического анализа, так как морфологическая принадлежность слова соответствовала его основным грамматическим значениям, без излишней детализации.

Особо следует отметить представление в НКРЯ такой неоднозначной части речи, как предикативы. Действительно, границы этого морфологического класса весьма дискуссионны. На основе проанализированных нами лексикографических ресурсов можно предположить наличие в современной лексикографии двух трактовок этого класса – широкой и узкой. Узкое понимание предикатива позволяет относить к этой части речи лишь некоторую группу слов с устойчивой семантикой состояния, сохраняющейся при употреблении слова как в связке со словом – носителем состояния, так и в качестве слова-предложения. В широком смысле к данному классу можно отнести также и большое количество слов, могущих в определенных контекстах выполнять роль оценочных предикативов, выражая субъективную позицию говорящего, оценку им ситуации. (Например, в предложении '*Бить детей антипедагогично*' можно было бы выделить предикатив '*антипедагогично*').

Ни первый, ни второй подход не удовлетворяют требованиям синтаксического анализа в полной мере. Действительно, в первом случае учитываются далеко не все слова, имеющие в современном русском языке устойчивое предикативное употребление; второй же подход, напротив, предполагает чрезмерное расширение класса предикативов, что затрудняет синтаксический анализ. НКРЯ удачно сочетает оба подхода, именно классификация этого ресурса используется авторами при уточнении множества предикатных слов.

Однако при решении проблемы уточнения лексикографических описаний существуют определенные трудности при работе с НКРЯ. В большой степени эти проблемы связаны с отсутствием возможностей

сортировки контекстов, распознавания тождественных контекстов. При выделении полного списка слов, причисляемых в НКРЯ к тому или иному морфологическому классу или подклассу, выводятся все контексты употребления (в том числе повторяющиеся) каждого из слов данного морфологического класса. Огромное количество во многом идентичных контекстов существенно затрудняет выявление всех представителей данной части речи.

Но в целом, положительный опыт использования НКРЯ в такой значимой проблеме синтаксического анализа, как разрешение функциональной омонимии, позволяет рассчитывать на успешное решение поставленных задач.

Следует отметить, что ресурсы НКРЯ могут применяться также в других методах разрешения омонимии. Так, весьма перспективным методом представляется составление для каждого омонима списка наиболее частотных разрешающих коллокаций. Для составления таких списков коллокаций целесообразно использовать текстовые коллекции Национального корпуса. Действительно, специфика сервиса поиска удобна для поиска коллокаций – например, поиск по морфологическим характеристикам. Особенности газетного корпуса позволяют выявить именно те коллокации, которые являются наиболее употребительными в настоящее время, тем самым упростить алгоритмы разрешения функциональной омонимии

Работа выполнена при финансовой поддержке РФФИ, грант № 05-07-90257.