УДК 81

ЛИНГВИСТИЧЕСКИЙ КОРПУС КАК ИНСТРУМЕНТ ИДЕНТИФИКАЦИИ ЗНАЧЕНИЯ ПОЛИСЕМАНТИЧНОГО СЛОВА

А.И. Доминикан

Военная Академия Воздушно-Космической Обороны
имени маршала Советского Союза Г.К. Жукова, Тверь

В статье рассматривается актуальная проблема лексической многозначности и пути её разрешения с опорой на контекст при помощи лингвистических корпусов. Приводятся определения терминов «лингвистический корпус», «полисемия» и «полисемантичное слово».

Ключевые слова: *семантика, корпусная лингвистика, контекст, многозначность, полисемия, разрешение многозначности.*

С приходом современных информационно-коммуникационных технологий в различные научные сферы, в частности и в лингвистику, возрастает популярность использования корпусов текстов для исследования различных аспектов языка. В данной статье мы рассматриваем лингвистический корпус в качестве инструмента для исследования значений полисемантичного слова.

Под термином «лингвистический корпус» принято понимать массив текстов, собранных в единую систему по определённым признакам (языку, жанру, времени создания текста, автору и т. п.) и снабжённых поисковой системой. Лингвистический корпус может включать как тексты литературных произведений, газет и журналов, так и транскрипты радио- и телепередач [10: 99]. На странице Википедии, посвящённой слову «корпус», лингвистический корпус определяется как «подобранная и обработанная по определённым правилам совокупность текстов, используемых в качестве базы для исследования языка. Они используются для статистического анализа и проверки статистических гипотез, подтверждения лингвистических правил в данном языке» [1].

В статье «Лингвистические корпуса: определение основных понятий и типология» [6] Н. В. Козлова, опираясь на данные немецкоязычного источника, приводит следующее определение: «Корпус представляет собой собрание письменных и устных высказываний. Данные корпусы, как правило, оцифровываются, то есть хранятся на компьютерах и доступны в электронном виде. При этом составные части корпуса, тексты, состоят из данных, а также, возможно, из метаданных, описывающих эти данные, и из лингвистических аннотаций, которые эти данные упорядочивают» (цит. по [6: 79]). В той же статье автор приводит достаточно подробное описание лингвистического корпуса в диахронии. Одним из наиболее интересных с нашей точки зрения

определений термина является следующее: «… a collection of *naturally*-occurring language text, chosen to characterize a state of variety of a language» [13: 171].

В данном определении речь идет о неотредактированных текстах, то есть язык представлен в том виде, в котором он проявил себя в речи, даже если это проявление является отклонением от языковой нормы [6: 80].

С это точки зрения самым большим корпусом можно считать Интернет (Web as Corpus), поскольку в сети в электронной форме и в свободном доступе представлено огромное разнообразие текстов. Но, как отмечают многие лингвисты, «тексты в Интернете представлены довольно хаотично, а лингвистически интересный запрос часто сложно или невозможно сформулировать с помощью языка запросов поисковой машины, по результатам поиска нельзя оценить представительность выборки и т.д.» [11: 128]. Здесь можно говорить о таком свойстве лингвистического корпуса текстов, как *наличие* или *отсутствие разметки*.

Разметка – это приписывание текстам и их компонентам специальных меток: внешних, экстралингвистических, структурных и собственно лингвистических, описывающих лексические, грамматические и прочие характеристики элементов текста [4: 6]. Метаразметка включает в себя сведения об авторе и о самом тексте [6: 81]. Разметка – это главная характеристика корпуса – она отличает корпус от простых коллекций текстов, в изобилии представленных в Интернете [11: 128].

Ещё одним свойством современного лингвистического корпуса выступает его *доступность в электронном виде*. При этом всё существующее множество корпусов текстов можно разделить на три категории:

- находящиеся в свободном доступе (например, Национальный корпус русского языка);
- находящиеся в частичном доступе (Британский национальный корпус);
- коммерческие (Bank of English) [6: 80].

Кроме того одной из важнейших характеристик корпуса текстов является *репрезентативность.* Так как корпус – это своеобразная модель языка, его репрезентативность определяет достоверность полученных на его основе данных. В корпусной лингвистике под репрезентативностью понимается сбалансированное и пропорциональное представление текстов в корпусе [9: 56]. Вопрос определения репрезентативности того или иного корпуса текстов является по сей день актуальным. Именно репрезентативность превращает обычный набор разнообразных текстов непосредственно в корпус текстов, пригодный для проведения лингвистического исследования. Репрезентативность лингвистического корпуса зависит от целей, которые ставит перед собой составитель корпуса: создание пользовательского корпуса либо национального корпуса определённого языка, претендующего на всеохватность языковых явлений, стилей, жанров и т. п. Вследствие этого

вопрос репрезентативности корпуса текстов является скорее вопросом из области объективности любого научного исследования [6: 81].

И, наконец, лингвистический корпус должен быть создан, несомненно, под определённую задачу, иными словами он должен быть *прагматически ориентированным*. Всё разнообразие существующих корпусов «определяется многообразием исследовательских и прикладных задач, для решения которых они создаются» [4: 12].

Говоря о типологии лингвистических корпусов, необходимо упомянуть наиболее часто выделяемые учёными типы корпусов. Например, по критерию языка представленных в корпусе текстов выделяют одноязычные, двуязычные и многоязычные корпусы. В свою очередь среди одноязычных корпусов можно выделить ещё две группы: корпусы, охватывающие весь язык и охватывающие только язык для специальных целей, например корпус медицинских текстов и т.д. [6: 82].

Кроме того выделяют устные, письменные и смешанные корпусы текстов. Как видно из названий, в устных корпусах представлены устные тексты, в том числе транскрипты теле- и радиопередач, выступлений и т.п.; в письменных – письменные тексты, например, художественные произведения, эссе, а в смешанных – оба типа текстов [там же].

Синхронные корпусы предполагают представление текстового материала для рассмотрения состояния языка как системы в определённый момент времени, тогда как для рассмотрения исторического развития какого-либо языкового явления либо всей языковой системы в целом существуют диахронные корпусы [там же].

И, как говорилось ранее, выделяют размеченные и неразмеченные корпусы.

Некоторые исследователи полагают, что корпусы текстов могут использоваться для разрешения лексической многозначности. Многозначность слова представляет собой актуальную лингвистическую проблему, так как зачастую возникают трудности при понимании текста, который переполнен полисемантичными словами.

Остановимся на понятии многозначности. Полисеми́я (от греч. πολυσημεία – ‘многозначность’) – многозначность, многовариантность, то есть наличие у слова (единицы языка, термина) двух и более значений, исторически обусловленных или взаимосвязанных по смыслу и происхождению [2].

В.В. Елисеева считает, что «многозначность, или полисемия слова – это наличие у языковой единицы более одного значения при условии семантической связи между ними или переноса общих либо смежных признаков или функций с одного денотата на другой» (цит. по: [3: 784]). В.Н. Немченко объясняет это явление «наличием у единицы языка более одного значения – двух или нескольких» [там же]. Таким образом, многозначное слово также называют *полисемичным* или *полисемантичным*.

А.Х. Мерзлякова в своей статье «Семантическая структура многозначных прилагательных» [7] подробно рассматривает структуру полисемантичного слова, опираясь на труды А.А. Уфимцевой, Ю.Д. Апресяна, Д.Н. Шмелева, И.А. Новикова, Ю. Найды и др. В данной статье автор делает обзор разных точек зрения как на статус значений, так и на отношения между значениями внутри многозначного слова. Согласно одной точке зрения все значения многозначного слова равноправны, но существует некое общее значение, так называемый «инвариант», по отношению к которому все другие значения выступают как его варианты. Другая – отвергает идею общего значения, но допускает наличие общего компонента у всех значений многозначного слова [7: 116].

Также существует мнение, что среди значений полисемантичного слова имеются денотативное (воспроизводимое) значение и производные от него значения (контекстуальные) [3: 784]. Контекстуальное значение слова влечет за собой вопрос о понятии «контекст».

Как правило, в речи слово употребляется в связи с другими словами, а не изолировано. Грубо говоря, контекст – это окружение слова. В словаре С.И. Ожегова приводится следующее значение данного термина: «

> КОНТЕКСТ, -а,м. (книжн.). Относительно законченная в смысловом отношении часть текста, высказывания. *Значение слова узнается в контексте*. || прил. контекстный, -ая, -ое и контекстовый, -ая, -ое» [8].

Под «языковым контекстом» обычно понимается «фрагмент текста или речи, содержащий избранное для анализа языковое выражение или единицу языка» или «ситуация употребления анализируемого выражения» [12: 1201]. Выделяются несколько уровней понимания контекста:

- минимальный контекст, в котором реализуются лексические и морфолого-синтаксические явления;
- текстовый контекст, включающий в себя фрагменты текста вплоть до текста целиком;
- контекст, предполагающий учёт текстов определенного типа (заданного функционального стиля, отобранной коллекции текстов и т.д.) (цит. по: [9: 57)].

Контекст не только выявляет данное значение слова, но и уточняет и конкретизирует его, создает вокруг него определённый круг ассоциаций [3: 789]. Некоторые учёные полагают, что контекст является *единственным* средством идентификации значения многозначного слова:

> «Context is the only means to identify the meaning of a polysemous word. Therefore, all work on sense disambiguation relies on the context of the target word to provide information to be used for its disambiguation» [13: 18].

Понятия «контекста» и «лингвистического корпуса» неразрывно связаны. Лингвистический корпус представляет практически неограниченные возможности для размежевания отдельных значений внутри многозначного

слова, так как языковые данные разного типа находятся в корпусе в своей естественной контекстной форме и позволяют исследователям уловить малейшие оттенки значений искомого слова. Огромные объёмы современных корпусов позволяют делать статистически значимые наблюдения о совместной встречаемости слов в разных значениях. При этом лучшие результаты дают аннотированные, или размеченные, корпусы [5: 90]. Создание больших аннотированных корпусов позволяет уточнять наборы значений слов не только в различных социальных группах или определённых предметных областях, но и в отдельных идиолектах. Эти данные также могли бы использоваться при разрешении неоднозначности (в условиях, когда отсутствует контекст, но известен говорящий) [цит. раб.: 98]. Кроме того корпусные исследования позволяют выявлять и уникальные авторские значения [цит. раб.: 99].

Подводя итоги, следует отметить, что лингвистический корпус является незаменимым инструментом для идентификации значения полисемантичного слова. Корпус текстов предоставляет контекст, который позволяет раскрыть малейшие оттенки значения исследуемого слова, а также нюансы его употребления. При этом необходимо подчеркнуть, что исследователь получает возможность работы с выборкой текстов, соответствующей целям его исследований.

Список литературы

1. Википедия [Электронный ресурс] / URL: https://ru.wikipedia.org/wiki/Корпус_текстов (дата обращения: 23.08.2018)
2. Википедия [Электронный ресурс] / URL: https://ru.wikipedia.org/wiki/Полисемия (дата обращения: 30.08.2018)
3. Галеева Т.И., Казиахмедова С.Х., Янова Е.А. Явление полисемии как феномен лингвистики // Вестник Удмуртского университета//История и филология. 2017. № 5. с.784–794
4. Захаров В.П. Корпусная лингвистика. СПб., 2005.
5. Иомдин Б.Л. Многозначные слова в контексте и вне контекста // Вопр. языкозн. 2014. № 4.С.87–103.
6. Козлова Н.В. Лингвистические корпуса: определение основных понятий и типология // Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация, 2013. Т. 11. № 1. С. 79–88.
7. Мерзлякова А.Х. Семантическая структура многозначных прилагательных // Вестник Удмуртского университета. Филологические науки, 2007. №5 (2). С. 115–122
8. Ожегов С.И., Шведова Н. Ю. Толковый словарь русского языка: 80 000 слов и фразеологических выражений / Российская академия наук. Институт русского языка им. В. В. Виноградова. 4-е изд., дополненное. М.: Азбуковник, 1999. 944 с.
9. Павельева Т.Ю. Изучение коллокаций на основе лингвистических корпусов текстов / Т.Ю. Павельева // Вестник Тамбовского университета. Серия: Гуманитарные науки. 2016. Т. 21. № 3–4 (155–156)) Тамбов, ТГУ, 2016. С. 56–61.
10. Сысоев П.В. Лингвистический корпус в методике обучения иностранным языкам // Язык и культура, 2010. №1 (9). С. 99–111: [Электронный ресурс] URL: http://www.lib.tsu.ru/mminfo/000349304/09/image/09-099.pdf (дата обращения: 18.12.2014).

11. Чернякова Т.А. Использование лингвистического корпуса в обучении иностранному языку // Язык и культура. 2011. № 4. С. 127–132.
12. Энциклопедия эпистемологии и философии науки / Ин-т философии РАН. Гл. ред. И.Т. Касавин. М.: Канон+, 2009. 1248 с.
13. N. Ide, J. Véronis. Introduction to the special issue on word sense disambiguation: the state of the art // Computational linguistics. 1998. V. 24. № 1 Pp. 2–40
14. Sinclair J. Corpus, Concordance, Collocation. Oxford, 1991. 179 p.

CORPUS AS A WORD SENSE DISAMBIGUATION TOOL

A.I. Dominikan
The Zhukov Military Aerospace Defense Academy, Tver

The work is devoted to a linguistic corpus as a word sense disambiguation tool. It gives a short overview of the following terms: «corpus», «polysemy» and «polysemous word»
***Keywords**: semantics, corpus linguistics, ambiguity, polysemy, context, WSD.*

Сведения об авторе:

ДОМИНИКАН Алина Игоревна – преподаватель кафедры иностранных языков Военной Академии Воздушно-Космической Обороны имени маршала Советского Союза Г.К. Жукова, e-mail: madara-san@list.ru