Наталья Анатольевна Прокофьева

Санкт-Петербургский государственный университет

n.prokofieva@spbu.ru

Екатерина Александровна Щеглова

Санкт-Петербургский государственный университет

e.scheglova@spbu.ru

**Методика лингвистической обработки текстовых данных, полученных автоматизированным путём**

Рассматривается необходимость лингвистической обработки данных, полученных в результате автоматизированного анализа. Объясняется необходимость совмещения методологии нескольких научных направлений при анализе текста. Недостаточность обращения к облачным структурам представления данных обосновывается обращением к научному опыту лексикографии.

Ключевые слова: автоматизированный анализ данных, лингво-семантический анализ, полевая структура языка, облачная структура.

Современность ставит перед исследователями новые вызовы: для представления релевантных результатов исследования требуется обработка всё большего количества данных (Big Data). Эту задачу легко решить при помощи машинной обработки текстов, более того, появляется возможность включения в эмпирическую базу текстов соцсетей и мессенджеров при условии, что исследователь задаёт адекватные условия поиска.

Но автоматизированный анализ данных не решает собственно научных вопросов, которые стоят перед исследователем. Подчас для получения подлинно ценных для науки результатов требуется совмещение методологических подходов нескольких научных направлений: например, социологии, политологии, культурологии, теории журналистики, лингвистики. Последнее представляется особенно важным, поскольку интерпретация текстовых данных невозможна без лингвистического инструментария. Из этого вытекает необходимость комбинирования машинной обработки эмпирической базы и собственно лингвистической аналитической работы. Скажем, простой частотный анализ не выявляет ключевых для текста, какого-либо дискурса или эпохи слов, тем более он не показывает стилевой принадлежности и стилистических оттенков, а также коннотаций, свойственных конкретному словоупотреблению в текущем моменте. Всё это становится возможным при дальнейшей ручной обработке текстового массива.

В большинстве случаев для представителей смежных научных направлений актуальна следующая информация о слове: ключевые слова текста/дискурса, семантика, сочетаемость, полевая структура. Полученные с помощью компьютерной обработки данные требуют подтверждения, обработки и последующего анализа специалистами. Так, для лингвиста очевидно, что на получении конкорданса работа только начинается: после этого идут этапы обработки полученных результатов в части составления лексико-семантических полей, определения синтагматических и парадигматических связей между словами. Этот анализ в конечном итоге и призван выявить ключевые единицы во всей совокупности их признаков.

Слово – подвижная единица, его семантика в каждом конкретном случае определяется совокупностью лингвистических и экстралингвистических факторов. Например, неизменно актуальная проблема определения границ лексического значения (разделения омонимии и полисемии) не может быть решена полностью автоматизированными методами и требует внимательного предварительного и последующего изучения контекстов употребления слова.

При определении сочетаемости слова автоматическая обработка данных не предполагает разделения собственно сочетаемости и состава предикатов, возможного для субъекта, выражаемого той или иной номинацией. Между тем для лингвиста очевидно, что это принципиально разные картины функционирования слова в дискурсе. Сочетаемость даёт представление об изменении семантики слова, движении внутри синтагматики и парадигматики, в конечном итоге – о системе оценочных смыслов, формируемых вокруг концептуально значимой для дискурса лексической единицы. Предикаты также могут указывать на оценочные смыслы, однако в большей степени формируют побудительность, а это уже следующая ступень осмысления обществом социально значимой ситуации.

Наконец, поля (семантические, лексико-семантические – в неспециализированных исследованиях их могут называть по-разному, не углубляясь в терминологическую разницу между этими понятиями) не могут формироваться простым набором слов, встречающихся в той или иной совокупности текстов. Поле имеет структуру (центр, ядро и периферию), которая строго определяет положение единицы в системе смыслов. Облачные структуры хороши своим иллюстративным потенциалом, но не дают полноценного представления об этой системе, а значит, менее информативны.

Строго говоря, лингвистика уже пережила этап безоговорочной веры в автоматизированные данные. Памятниками той эпохи остались словари Ю. Н. Караулова [1; 2], которые, дав ценные сведения для дальнейших исследований, стали подтверждением невозможности исключения человека при обработке языковых данных.

Литература

1. Караулов Ю. Н. Частотный словарь семантических множителей русского языка. М., 1980.

2. Караулов Ю. Н., Молчанов В. И., Афанасьев В. А., Михалев Н. В. Русский семантический словарь: опыт автоматического построения тезауруса: от понятия к слову. М., 1982.