Людмила Рычкова

Лингвистические базы данных как материал и средство исследования и описания языковых изменений

Studia Rossica Posnaniensia 28, 79-84

1998

STUDIA ROSSICA POSNANIENSIA, vol. XXVIII: 1998, pp. 79-84. ISBN 83-232-0887-5. ISSN 0081-6884.
Adam Mickiewicz University Press, Poznań

ЛИНГВИСТИЧЕСКИЕ БАЗЫ ДАННЫХ КАК МАТЕРИАЛ И СРЕДСТВО ИССЛЕДОВАНИЯ И ОПИСАНИЯ ЯЗЫКОВЫХ ИЗМЕНЕНИЙ*

LINGUISTIC DATA BASE AS A MATERIAL AND A MEANS IN STUDYING AND DESCRIPTION OF LANGUAGE CHANGES

ЛЮДМИЛА РЫЧКОВА

ABSTRACT. On the example of special computer text corpora linguistic data base is regarded as a material and a tool for language changes studying and description.

Людмила Рычкова, Гродненский государственный университет им. Янки Купалы, Кафедра общего и славянского языкознания, Лаборатория теоретической и прикладной лингвистики, ул. Ожешко 22, 230023 Гродно, Беларусь.

Необходимость автоматизации лингвистических исследований сегодня является очевидной и обусловливается, наряду с повышением степени объективности результатов исследований, возможностью получения совершенно новой информации о языковом материале, получить которую иным путем не представляется возможным. Отсюда – актуальность проблемы формирования лингвистических баз данных, прежде всего основанных на больших компьютерных файлах текстов. Такие „текстовые" базы данных могут рассматриваться как специальным образом организованный материал исследования и, одновременно, служить средством описания языковых изменений.

Методика формирования лингвистических баз данных вообще и текстовых в частности находится в стадии становления. Очевидно, что объективно она должна базироваться на потенциальных направлениях анализа, сложившихся в рамках лингвистических представлений о тексте.

Термин „лингвистика текста" давно и прочно закрепился в языкознании, отразив тем самым проявление нового объекта лингвистических исследований наряду с традиционными „предложением (высказыванием)", „словом (словоупотреблением)", „морфемой (морфом)". Два основ-

* Здесь представлены результаты работы, выполненной в рамках научно-исследовательской темы, финансируемой Министерством образования Республики Беларусь.

ных „признака", характеризующих новый объект анализа, можно определить как „связность" и „цельность"[1]. Оба этих признака, как и сам текст, являются двуплановыми: они функциональны и имеют определенный набор формальных показателей, которые можно описать и до некоторой степени точности исчислить.

Став самостоятельным объектом исследования, текст не перестал быть „языковым материалом" (в терминологии Л.В. Щербы), который служит целям исследования других объектов (любых составляющих его линейных единиц).

Наконец, текст остается объектом исследования традиционной текстологии, изучающей языковые средства, отражающие композиционно-стилистические особенности жанра, творческую манеру писателя, стиль автора.

Различные аспекты рассмотрения текста определяют потенциальные возможности направления его анализа. Так, текст может рассматриваться как сегментная единица, имеющая начало, конец и протяженность, измеряемую в иных, более коротких, сегментных единицах. Очень важное для текста понятие границ сопряжено с его объемом, то есть протяженностью, и определяется через отсутствие связи между текстом и последующими предложениями[2].

Различия в целях коммуникации ведут к выбору определенных структурных закономерностей организации текста, его конструирования, которые могут быть сведены к понятию „типа" или „жанра" текста. В зависимости от степени „жесткости" структуры текста в нем могут преобладать либо общие, либо индивидуальные черты. Доминирование последних „присуще прежде всего художественным текстам и связано отчасти с «мягкостью», открытостью их структуры"[3].

Текст может анализироваться с точки зрения закономерностей построения и сочетаемости составляющих его единиц, а также с точки зрения тех языковых приемов и средств, которые определяют стиль либо „наиболее интересны у данного автора и в данном произведении", наиболее значимы „для раскрытия содержания"[4].

„В аспекте учета участников речевой коммуникации" тексты могут рассматриваться „как диалоги, монологи, тексты смешанного типа, полилоги"[5].

---

[1] Ю.В. Попов, Т.П. Трегубович, *Текст. Структура и семантика*, Минск 1984, с. 190.

[2] М.И. Откупщикова, *Синтаксис связного текста*, Ленинград 1982, с. 30.

[3] Ю.В. Ванников, *К обоснованию общей типологии текстов, функционирующих в сфере научно-технического перевода.* В: *Текст как объект лингвистического анализа и перевода*, Москва 1984, с. 18.

[4] М.П. Демидова, Г.Н. Моложай, *Лингвистический анализ текста*, Минск 1988, с. 3, 27.

[5] М.И. Окупщикова, указ. соч., с. 8.

Текст, „созданный конкретным автором, в конкретное время, в конкретной ситуации"[6], является объектом стилеметрии, прикладной филологической дисциплины, ориентированной на измерение стилевых характеристик текста, включая его индивидуально-авторские особенности. Изучением последних занимается диагностика, основной задачей которой является „оптимальная свертка исходного признакового пространства художественного текста с целью выделения наиболее информативных диагностирующих признаков"[7], каковыми, без сомнения, и являются индивидуально-авторские особенности текста.

Конкретные „специфические признаки стиля, слога, словоупотребления"[8] пока еще не выделены и недостаточно изучены. Однако поскольку современная лингвистика рассматривает стиль прежде всего как категорию структурно-семантическую, то очевидно, что центральную роль в стилеметрии играет именно синтаксический уровень: „В синтаксическом строе наиболее откровенно проявляется авторская манера развертывания мысли, которая может быть проинтерпретирована в „содержательных" терминах, таких, как слитность – разорванность мысли, ее синтетичность – аналитичность, напряженность – уравновешенность, простота – сложность, предметность – процессуальность и т.д."[9].

При изучении специфических черт индивидуально-авторского стиля текста в априорный набор анализируемых параметров рекомендуется включать „первичные параметры, значения которых определяются непосредственно" в тексте[10]. Набор таких параметров может быть разным для каждого автора и каждого произведения.

Возможность автоматизации тех или иных направлений лингвистического анализа текста обусловливается степенью эксплицитности признаков, которые могут быть положены в основу такого анализа. Так, наибольшей степенью „автоматизуемости" будет обладать любой вид анализа, который исходит из формы. Например, наличие в письменном тексте естественных делимитаторов, таких как пробелы между словами или знаки препинания, обозначающие конец предложения, позволяет без труда автоматически сегментировать текст, однако такое сегментирование не учитывает единства формы и содержания языковых знаков.

Изначальное задание формальных характеристик типа текстемы делает возможным автоматическое отнесение текста к определенному жанру,

---

[6] Г.Я. Мартыненко, С.В. Чебанов, *Стилеметрия*. В: *Прикладное языкознание*, Санкт-Петербург 1996, с. 422-425.

[7] Там же.

[8] М.А. Марусенко, *Атрибуция анонимных и псевдонимных текстов методами прикладной лингвистики*. В: *Прикладное языкознание*, указ. соч., с. 469-473.

[9] Г.Я. Мартыненко, *Сложность синтаксических структур и стилистическая диагностика*. В: *Прикладное языкознание*, указ. соч., с. 435-436.

[10] М.А. Марусенко, указ. соч.

а задание какого-либо специального формального признака позволяет автоматически выделять любые микротекстовые структуры, обладающие этим признаком.

В любом случае результаты анализа, который исходит только из формы, будут иметь ту или иную степень погрешности, допустимость которой определяется целями анализа. Если целью анализа является изучение некоторых семантических аспектов, но текст не представляет возможности их формального изначального определения, то в этом случае для автоматизации анализа необходимо ввести в исходный текст специальные индексы – показатели „скрытой" семантики. Индексы могут присваиваться автоматически (в случае, если допустимы значительные погрешности индексации), вручную (что существенно повышает точность индексации и, следовательно, результатов анализа), либо сочетанием двух способов.

Выделение потенциальных направлений анализа определенного текстового корпуса – необходимое условие конкретного формирования компьютерных текстовых файлов, предназначенных для целей многоаспектного лингвистического анализа этого корпуса. Сама по себе компьютерная форма существования текста (компьютерная версия) не дает особых преимуществ в проведении исследований по сравнению с традиционными формами его представления. Такие преимущества дают, однако, текстовые базы данных, среди которых особое место занимают „полнотекстовые" (в иной терминологии – „цельнотекстовые") базы. Особенности формирования таких баз данных, их использования в качестве материала лингвистических исследований и основы автоматического лингвистического анализа достаточно подробно описаны нами ранее[11]. В зависимости от системы, принятой в базе индексации, выделяют общецелевые текстовые базы данных и базы данных специального назначения[12].

Для общецелевых цельнотекстовых баз данных характерна система индексации, обусловленная лишь задачами структуризации текстов и разграничения нетождественных словоупотреблений. Существенно увеличивает спектр потенциальных направлений анализа расширение системы индексации, цель которой – эксплицитное выражение первичных параметров анализа в компьютерном текстовом файле. Сам процесс априорного выделения как самих параметров анализа, так и их признаков, под-

---

[11] См., например: Л. Р ы ч к о в а, *Полнотекстовые базы данных как материал для лингвистических исследований.* В: *Е. Карский и современное языкознание. Материалы шестых научных чтений*, ч. 1, Гродно 1996, с. 227-230; Л.В. Р ы ч к о в а, *Цельнотекстовые базы данных как основа автоматического лингвистического анализа.* В: *Tekstas ir Kontekstas. Тезисы докладов научной конференции*, Шауляй 1996, с. 99-100.

[12] У.Н. Ф р э н с и с, *Проблемы формирования и машинного представления большого корпуса текстов.* В: *Проблемы и методы лексикографии*, „Новое в зарубежной лингвистике", Москва 1983, вып. XIV, с. 334-352.

лежащих индексации, достаточно сложен. Во-первых, среди признаков могут быть „шумовые"[13], то есть такие признаки, выделение которых не только является избыточным, но и ведет к понижению точности анализа. Во-вторых, увеличение системы индексов ведет к „засорению" базы и усложняет компьютерную обработку. В-третьих, необходимость внесения в текст множества помет неизбежно ведет к ошибкам ввода и требует весьма затруднительной и дорогостоящей процедуры проверки сформированных компьютерных файлов.

Если параметры анализа рассматривать с точки зрения анализируемых объектов, а признаками параметров, требующими формального выражения в текстовой базе данных, считать признаки этих объектов, то такой текстовый файл может рассматриваться как информационно-поисковый массив, в котором может осуществляться поиск любых текстовых единиц – объектов, обладающих определенными значениями признаков, наборов признаков и их комбинаций.

Формирование любого информационно-поискового массива всегда опирается на прагматический аспект, связанный с информационными потребностями, которым должна удовлетворять система. Определить эти потребности можно, исчислив типы потенциальных запросов к системе. Среди всех запросов можно выделить типовые, ответ на которые можно получить исходя из возможностей цельнотекстовой базы данных общецелевого назначения, и специальные, для осуществления поиска по которым нужна дополнительная индексация.

Специфичным видом типовых запросов можно считать „лексикографические" направления анализа[14]: поиск местонахождения любого слова в тексте, показ его в контексте с указанием точного места употребления в тексте-оригинале. Речь идет о построении различных словарей-словоуказателей, включая конкордансы, что стало уже делом достаточно обычным, как и построение различных видов частотных словников.

Основываясь только на формально выраженных признаках структуризации текста, можно совершенно определенно выделить ряд „взаимовложенных" линейных сегментных объектов: текст, реплика, предложение, словоупотребление.

Наличие обязательной индексации в текстовой базе общецелевого назначения позволяет существенно расширить набор признаков поиска, а за счет их комбинации многократно увеличить возможности системы. Перечень типовых запросов к такой системе может включать (поиск может производиться по всем текстам, отдельному тексту, в пределах авторских ремарок, в пределах каждой реплики каждого персонажа и в любой

---

[13] М.А. М а р у с е н к о, указ. соч.

[14] D. B i b e r, *Applied linguistics and computer applications.* В: *Introduction to applied linguistics. Reading*, MA: Addison-Wesley 1992, с. 257-278; *Advanced computing in the humanities*, Bergen 1996, с. 20.

их совокупности, а результаты выдачи могут учитывать необходимость сопровождения контекстом определенной длины):

- нахождение словоупотреблений (с определенным признаком либо комбинацией признаков);
- нахождение различных типов омонимов (либо любого их подмножества, характеризующегося определенными признаками);
- нахождение определенных сочетаний словоформ (с учетом признаков длины, частеречной принадлежности, омонимии либо различных комбинаций этих признаков);
- нахождение предложений, характеризующихся определенной пунктуационной оформленностью (можно определенной длины и/или характеризующихся определенным набором частей речи, можно с учетом их распределения);
- нахождение реплик/текстов, характеризующихся определенным признаком (их набором, комбинацией).

Введение дополнительного признака имени собственного, служащего также цели разграничения тождественно оформленных нетождественных объектов, и дополнительного признака для формально неоднословных языковых знаков позволяет не только существенно расширить спектр возможных запросов к системе, но и повысить степень точности лингвистического анализа.

Правильное построение и тестирование модели базы данных обеспечивает очень быстрый поиск по различным критериям. Сама база данных может содержать большое количество различных таблиц, связанных между собой определенными условиями. Каждая таблица содержит только те данные, которые соответствуют ее функциональному назначению. Например, в таблице индексов хранится только уникальный индекс словоупотребления (индекс частеречной принадлежности и индекс омонимии, если он имеется) и ссылка на следующую таблицу, в которой хранится иной вид информации, позволяющий идентифицировать словоупотребление в тексте. Все таблицы в базе данных связаны между собой правилом „подчиненный-главный", что позволяет использовать для выборок наиболее эффективный язык запросов и добиться высокой производительности.