КРИТИКА И БИБЛИОГРАФИЯ

А.-Л. Согваль

«Система автоматического анализа словоформ (применительно к русскому языку)»

Книга Анны-Лены Согваль «Система автоматического анализа словоформ (применительно к русскому языку)» [1], вышедшая в серии книг по математической лингвистике, привлекает внимание преподавателей, методистов и исследователей прежде всего тем, что она дает конкретные рекомендации по применению машинного анализа русского текста в прикладных, учебных, целях. Программа для описанного в книге анализа словоформ составлена таким образом, что, как об этом говорит сам автор, могут быть получены следующие данные: распределенность и богатство словаря проанализированного текста; сравнение употребительности конкретного слова в тексте с его положением в существующих частотных словарях; сравнение словарей разных текстов для определения порядка изучения отдельных слов и скорости, с которой может быть сформирован словарь студента путем чтения таких текстов; для неродственных языков, пример которых может дать сопоставление шведского и русского, определение степени градуированности словаря иностранного языка для начального этапа обучения; определение состава словоформ в тексте и их распределенность и употребительность; присутствие и качество грамматических вариантов в исследуемых текстах (стр. 19).

Результаты анализа были использованы при составлении учебных материалов по русскому языку (подбор текстов для чтения) в Славянском институте Упсальского университета.

Эти результаты интересны и ценны сами по себе. Многостороннее исследование лексического состава текстов и описание состава словоформ каждого слова в тексте дают надежную основу для преподавателей, авторов учебников, учебных пособий, серий книг для чтения; для методистов, создающих лексические и грамматические минимумы разного характера; для составителей тематических списков лексики; для авторов разнообразных учебных словарей.

Ценность и надежность анализа — прямое следствие продуманной и научно обоснованной программы анализа. Методика разработки программы машинного анализа словоформ русского текста и составляет основное содержание книги А.-Л. Согваль. Эта единая методика разработана в двух направлениях: составление системы алгоритмов для анализа словоформ — «грамматика» — и составление словаря исследуемого текста (исследуемых текстов).

В книге можно найти подробное описание программ анализа словоформ и составления словаря, а также методики работы и последовательности операций собственно машинного анализа на ЭВМ-370/155, т. е. эта книга может быть использована как инструкция при составлении подобных программ для машин определенного класса. Эта программа составлена с учетом основных публикаций, относящихся к области автоматического анализа русского текста и машинного перевода [2]. Но особенно интересны приемы и методы лингвистического анализа словоформ русского печатного текста, поскольку, и это уже неоднократно замечалось исследователями, разработка системы машинной обработки текстов нередко дает прямые выходы в практику преподавания иностранных языков, в том числе русского как иностранного.

Задача, стоящая перед автором книги,— создать программу для машинной обработки русского печатного текста — исключала возможность обращения к синтаксическим и семантическим признакам слова в тексте и заставляла ограничиваться чисто формальными признаками при анализе флективной системы русского языка. Однако автору удалось разработать такую программу анализа, в которой были учтены все формально-грамматические характеристики слов, отраженные во флективной системе русского языка.

При составлении алгоритмов (системы правил обработки слов текста и порядка применения этих правил в процедуре обработки) А.-Л. Согваль делит все русские слова на классы в зависимости от набора и качества флексий (или отсутствия флексии у слова), принимаемых словом в системе русского языка. Под «качеством» флексии понимается и способность флексии указывать на одно или несколько грамматических значений и свойство грамматического значения, например, грамматическое значение рода у существительных — классифицирующая

[1] Anna-Lena Sågvall. A System for Automatic Inflectional Analysis (Implemented for Russian). Almqvist and Wiksell, Stockholm, 1973.

[2] См. библиографию на 141—142 стр. книги.

категория, у прилагательных — согласуемая («парадигматическая» в терминологии автора). Все слова русского языка распределены по 7 классам. Первый класс (класс формальных существительных) — слова, флексия которых указывает на число и падеж как парадигматические категории и на род как классифицирующую («наследственную» в терминологии автора) категорию (стр. 21). Третий класс объединяет все слова, у которых род, число и падеж — парадигматические, т. е. согласуемые категории; слова этого класса обладают полными местоименными окончаниями (стр. 21—22); наличие классифицирующей категории — степени сравнения — становится основанием деления членов этого класса на два подкласса. Второй класс включает тоже согласуемые слова, но обладающие смешанными — полными и краткими местоименными окончаниями типа *он*, *наш*, *кошкин*. В четвертый класс помещены слова, флексии которых указывают только на категорию падежа как парадигматическую категорию — собственно-личные местоимения *я*, *ты*, *мы*, *вы*, вопросительно-относительные местоимения *кто*, *что*, возвратное местоимение *себя* и количественные числительные от 3 и далее (стр. 22). К пятому классу относятся слова *два*, *оба*, *полтора*, флексии которых указывают на парадигматические категории рода и падежа. Шестой класс — все неизменяемые слова русского языка: предлоги, союзы, наречия, частицы, вводные слова (стр. 22). Седьмой класс — глаголы, окончания которых выражают парадигматические значения времени, лица, числа, залога, рода, падежа (для причастий). При описании свойств флексий слов, входящих в седьмой класс, делается замечание, что у слов этого класса, как и у слов других классов, во флексии может быть выражен не весь набор возможных значений. Подобная возможность предусматривается процедурой анализа. Это последнее замечание представляется важным, так как при определении слов по выделенным классам автором были «проведены некоторые модификации слов анализируемых текстов, преследующие цель сокращения числа лексем для экономии анализа» (стр. 20). При этом произошли неизбежные потери точности анализа, о чем говорит и сам автор. Такие досадные потери, на наш взгляд, относятся в первую очередь к отказу от целой части речи в русском языке — категории состояния. Возможность многих членов этой части речи быть отнесенными по своим формальным признакам к существительным — *жаль*, *пора* (однако каждое из этих «существительных» будет в любом тексте представлено только одной словоформой, т. е. должно быть признано неизменяемым словом), прилагательным — *рад*, *готов*, *должен*, наречиям — *стыдно*, *охотно* — действительно упрощает анализ, но при этом происходит потеря качества системы русского языка. Тем более что краткие прилагательные переводятся в разряд полных — *должен* —*должны* (стр. 20), так как, по мнению автора, они стоят в одном ряду с *хорош* — *хороший*. В класс прилагательных помещаются и наречия на **-о**, исторически образованные от прилагательных *вероятно* — *вероятный* (стр. 20). Но в современном русском языке нет существительного *жаль*, прилагательных *охотный* (в названии улицы «Охотный ряд» сохраняется архаическая форма, не связываемая носителями языка с «охота»), *радый*, *стыдный*. Пары *должен* и *должный*, *готов* и *готовый* и под. так разошлись в своих значениях, что целесообразнее рассматривать их как разные слова, а не как разные значения одного слова[3]. По своим чисто формальным характеристикам в современном русском языке (в силу своей синтаксической роли) слова из категории состояния — краткие прилагательные по происхождению (*был/будет*) *должен*, *рад*, *готов* и под.— абсолютно равны кратким страдательным причастиям типа (*был/будет*) *открыт*, *сделан* и под. Представляется также целесообразным выделение восьмого класса — класса категории состояния, родственного седьмому классу — классу глагола, подобно тому как среди слов с именными флексиями были выделены третий и пятый классы. В этот класс вошли бы те слова категории состояния, которые представлены одной словоформой типа (*было/будет*) *жаль*, *надо*, *можно* и под.

В практике обучения русскому языку как иностранному есть попытки, опираясь на фортунатовскую формальную школу, представить систему частей речи в русском языке с позиций поведения слова в составе предложения и формального выражения этого поведения в окончаниях слова. С этой точки зрения все слова русского языка делятся на изменяемые (имеющие набор флексий) и неизменяемые. На следующем этапе разделяются слова с именными окончаниями — имена и глагольными окончаниями — глаголы; затем следует деление всех имен на имена с независимой парадигмой (имена существительные, количественные числительные, собственно личные местоимения) и на имена с зависимой парадигмой (все согласуемые части речи русского языка)[4].

На последующей стадии обработки для каждого класса слов, кроме шестого, составляется набор флексий, при этом разрабатываются особая система и процедуры, определенные «шаблоны», «выравнивающие» графическое оформление флексий у слов с мягкой и твердой основой (стр. 27), «устраняющие» беглые гласные в корне и основе (стр. 31), учитывающие корневые и позиционные чередования букв (звуков) в составе одной лексемы (стр. 32-33, 45—46). Наряду с этим вырабатываются соответствующая система и процедуры, ведущие к снятию омографов на разных этапах анализа, начиная от различения слов, например союза *потóм* и твор. пад. ед. ч. существительного *пóтом* от *пот*, и кончая различением форм слова, например *ноге* (дат. пад.) — (*в*) *ноге* (предл. пад.), *узнаю* — *узнáю*.

Здесь интересно отметить, что значимые для машинного анализа характеристики слова оказываются равными единицам обучения, выделяемым в практике преподавания русского языка как иностранного на начальном этапе обучения[5].

Составление программы анализа словоформ проводится вместе с разработкой программы составления словаря. Стадии и техника составления словаря обобщены автором в схеме, воспроизводящей механизм работы машины.

В заключение представляется необходимым отметить еще одну практическую ценность работы. В качестве результата обработки текста получаются не только алфавитные списки словоформ текста, но каждая словоформа сопровождается набором контекстов (предложений) (см. примеры на стр. 138). Трудно переоценить возможности подобной картотеки, позволяющей исследователю обращаться к жизни словоформы в тексте.

Е. М. Степанова

[3] См., например: «Грамматика современного русского литературного языка». М., 1970; С. И. Ожегов. Словарь русского языка. М., 1952, стр. 120, 147—148, 309.

[4] «Программа. Практический курс русского языка для филологических факультетов высших учебных заведений». М., 1971, стр. 13—37; «Программа по русскому языку для учащихся зарубежных вузов естественно-технического профиля». М., 1972, стр. 48—63.

[5] «Программа по русскому языку для курсов и кружков». М., 1971, стр. 66—74.