Н. П. Дарчук

КОМПЬЮТЕРНАЯ ГРАММАТИКА УКРАИНСКОГО ЯЗЫКА *АГАТ* И ПЕРСПЕКТИВЫ ЕЕ ИСПОЛЬЗОВАНИЯ

Потребность в лингвистическом обеспечении систем «человек – машина – человек» обусловлена необходимостью удобной кооперации человека и машины, опирающейся на естественный язык. В социальном плане значимость лингвистических проблем компьютеризации связана с такими основными направлениями индустрии обработки знаний, как сбор, хранение, систематизация, распространение, интерпретация информации, для чего создается специальное лингвистическое обеспечение.

Лингвистическое обеспечение автоматических систем – совокупность средств для осуществления компьютеризации языковой деятельности – необходимо практически для любой интеллектуальной деятельности человека. С технологической точки зрения речь идет о создании того или иного типа автоматической системы обработки информации, на входе и выходе которой текстовая информация на естественном языке. Типы систем разно-

образны и направлены на моделирование разных языковых задач, в частности, таких как диалоговое взаимодействие, сжатие информации, реферирование текста, логическая обработка содержания, перевод и т.д. [1; 2; 3]. Прикладные системы, которые создает лингвист в этой области, – это **лингвистически осмысленные метаязыки** – модели представления знаний, каждая из которых основывается на **теоретических положениях языкознания** и реализуется с помощью методов **структурно-математической лингвистики**. Воплощенная в прикладные задачи диалектическая трихотомия «традиционная лингвистика – структурно-математическая лингвистика – компьютерная лингвистика» способствует высокому уровню лингвистического обеспечения автоматических систем.

Для автоматического анализа украинскоязычного текста сотрудниками лаборатории компьютерной лингвистики Института филологии Киевского национального университета имени Тараса Шевченко создана компьютерная грамматика АГАТ. **Компьютерная грамматика** – это системное, строго упорядоченное, формализованнное, лингвостатистическое, интегральное описание знаковых единиц определенного языка в виде структурных моделей с необходимой и достаточной аналитикой для выполнения задач искусственного интеллекта, которые воспроизводят и имитируют исследовательскую деятельность лингвиста [4]. Компьютерная грамматика АГАТ имеет такие особенности: **уровневый подход** (уровни взаимодействуют между собой от нижнего до верхнего, каждый следующий уровень использует результаты анализа предыдущего); **открытость** стратификационной структуры грамматики, что является принципиальным моментом, поскольку позволяет довольно точно расширять объем лингвистического обеспечения, усложнять словарное и модифицировать программное обеспечение без перестройки всей системы, которая является иерархическим комплексом компьютерных моделей: морфемно-словообразовательной, морфологической, синтаксической моделей, построенных на основе формальных, точных и однозначных правил. Эти модели можно считать исследовательскими, потому что заложенные в грамматики алгоритмические правила приводят к выявлению того или иного языкового явления (морфов, словоформ с их частеречными и категориальными характеристиками, словосочетаний, деревьев зависимостей предложений и т.п.). Алгоритмически сымитирована деятельность лингвиста – а именно обеспечен переход от совокупности текстов к системе, которая лежит в их основе, выявлены элементарные единицы и классы элементарных единиц. Разработанные модели являются моделями анализа, индуктивными, несемантическими и детерминистскими (структурными).

Язык – относительно открытая система относительно открытых подсистем, каждую из которых можно моделировать и устанавливать определенные отношения между подсистемами. Как свидетельствует практика, большинство современных лингвистических процессоров модульного типа, адекватных поэтапному членению процесса анализу, отвечает уровням языка: фонемному (единица фонема/аллофон), морфемному (морфема/морф), лексическому (лексема/словоформа), синтаксическому (модели словосочетаний, модели предложений). Семантический модуль не имеет уровневого

соответствия в языковой системе, но он на семном уровне завершает (на данном этапе разработки) смысловой анализ украинского текста. Все единицы модулей взаимодействуют (АГАТ отражает их последовательную работу), но морфологические, синтаксические, семантические характеристики обрабатываются разными алгоритмами и программами с разным словарным обеспечением. Системно адекватным этому является такое лингвистическое обеспечение, которое имеет открытый модульный характер, и в некоторых случаях предусматривается их совмещенность для пополнения данных и коррекции результатов (например, после автоматического морфологического анализа обоих этапов остается несколько процентов словоформ с неснятой омонимией, которая окончательно снимается на синтаксическом уровне). Система может выполнять отдельные конкретные задачи анализа (напр., только морфологического). При этом открытость является главным фактором эффективного функционирования АОТ, гарантией того, что для введения новой информации не потребуется перестраивать всю систему.

Словарное обеспечение является информационной основой модели и реализуется в виде иерархической, модульно совместимой и открытой системы. Компоненты, которые представляют лингвистическую модель АОТ, это лингвистические процессоры, которые последовательно, один за другим обрабатывают входной текст. Вход одного процессора – это выход из другого. В созданной системе выделяются такие модули:

- Морфологический анализ. Построение морфологической аннотации слов входного текста;
- Синтаксический анализ. Выделение словосочетаний. Построение дерева зависимостей всего предложения;
- Морфемный анализ. Членение входного текста на морфы;
- Семантический анализ. Построение тезауруса текста.

Для каждого уровня разработан метаязык его представления – константы и правила их комбинации. На морфологическом уровне – константами являются граммемы (род, число, падеж, время, лицо, наклонение), на морфемном – тип морфа, на синтаксическом – тип словосочетания, тип связи, на семантическом – семантические категории. Важным фактором является поступательность в работе модулей: следующий анализатор улучшает результаты предыдущего уровня. Например, на синтаксическом уровне «доснимается» грамматическая и лексико-грамматическая омонимия, а семантический анализатор помогает достроить синтаксический граф.

Обязательной составляющей компьютерной грамматики является **автоматический морфологический анализ** (АМА) словоформ, потому что ни морфемный, ни синтаксический, ни семантический анализы не могут обойтись без определения для словоформы ее частеречной характеристики и словоизменительных форм. В **задачи АГАТ-морфологии входят**: автоматическое определение для единиц текста грамматической информации о месте их в морфологической системе языка; автоматическая идентификация словоформ одной лексемы. Морфологическая аксиоматика была налажена на возможность алгоритмического оперирования грамматическими данными. В прикладном аспекте создан словарь квазиоснов объемом в 210 тыс. оди-

ниц и, соответственно, словарь словоформ, которые порождаются соединением информации, взятой из таблицы основ и вспомогательной таблицы, – приблизительно 3,2 млн. словоупотреблений, что обеспечивает приписывание морфологической информации словоформам практически на 97 %. Методологически АГАТ-морфология украинского языка является автоматическим формально-морфологическим процессором с элементами морфолого-синтаксического анализа. Особое внимание при создании АГАТ-морфологии уделено определению речевых условий, в которых реализуются актуализированные грамматические значения единицы-омонима. Был создан Грамматический словарь омоформ, а также определены речевые условия для реализации значений исследуемой словоформы, сформулированные с помощью лингвистического метода – контекстного анализа (КА). В основе КА лежит утверждение о том, что многозначные элементы языка функционируют в своих конкретных значениях в определенном лексико-грамматическом контексте. Реализация этой идеи нашла отображение в создании автоматического конкорданса, теоретической основой которого являются:

1) наличие таких определителей, по которым каждое значение словоформы (грамматическое, лексическое) детерминируется в контексте другими словоформами, их сочетаниями или другими текстовыми признаками;

2) текстоцентрический подход к его созданию: он составляется на определенном массиве текстов для определенной словоформы или лексемы. Такой словарь-конкорданс исчерпывающе иллюстрирует использование определенной лексемы и всех ее ЛСВ с лексико-грамматическими значениями.

АГАТ-синтаксис компьютерной грамматики украинского языка создавался как лингвистический процессор, настроенный на моделирование синтаксической структуры входного предложения на уровне словосочетаний (1-й этап) и дерева зависимостей (2-й этап). Результат анализа – синтаксическая структура предложения, которая является совокупностью данных о синтаксических связях слов / словоформ в словосочетании – минимальной единице предложения.

АГАТ-синтаксис базируется на **формально-синтаксической теории** представления предложения. Это комплекс алгоритмических операций, которые выполняются над цепочками информации морфологического характера, представленными в исходном тексте, для установления синтаксических связей между текстовыми единицами. Практическая реализация теоретических положений осуществлялась путем взаимодействия двух структурных методов: для представления синтаксической структуры предложения в терминах словосочетаний применен **метод непосредственно составляющих**, а структуры целого предложения – **дерево зависимостей**. Алгоритмически и программно в синтаксическом модуле можно осуществлять **переход от непосредственно составляющих к дереву зависимостей:** корнем дерева является глагол-сказуемое, в узлах предложения находятся словоформы, каждая дуга дерева, которая связывает пару узлов, интерпретируется как подчинительная связь.

Значение АГАТ-синтаксиса заключается в том, что, опираясь на теоретический синтаксис в делении словосочетаний на и**менные, адъективные,**

местоименные, глагольные и наречные, можно автоматически выявлять тип сочетаемости – **предикативный, подчинительный и сочинительный** – каждого полнозначного слова в тексте. В соответствии с концепцией АСА при выделении словосочетаний предусматривался предварительный этап – создание грамматики валентностей с подграмматиками для глагола, существительного, адъектива, а также словника фразеологизмов и коллокаций. Созданием такой информационной базы в виде подграмматик валентностей **были расширены возможности украинского теоретического синтаксиса в получении** из текстов разного стилевого и жанрового направления **информации** о конструктивных возможностях сочетаемости каждой части речи и типовых моделях для определенной части речи. Установление по правилам АГАТ-синтаксиса для каждого слова **подчинительных, предикативных и сочинительных** типов связей является воспроизведением общей системы отношений между компонентами описываемой ситуации в предложении. Перспективу АГАТ-синтаксиса мы видим в соединении его с семантикой, точнее – в соединении лексики и грамматики, поскольку влияние семантики на сочетаемость общепризнана и виды синтаксической связи между связанными словами являются производными от их семантики.

Разработка АГАТ-синтаксиса в пределах АОТ связана с общетеоретической необходимостью изучения сочетаемости лексических единиц, что открывает возможность в современной украинистике исследовать грамматическую и лексическую валентность слов, моделировать типовую сочетаемость класов слов, синонимию словосочетаний разных структурных типов, опираться на лексическую и грамматическую валентность как критерий синонимичности, изучать законы комбинаторики словосочетаний разных типов и разрядов или разграничение свободных и фразеологических словосочетаний, взаимодействие стойкости и идиоматичности и пр. Несмотря на то, что перечисленные проблемы так или иначе рассматриваются в теоретической грамматике, АГАТ-грамматика открывает новые перспективы в **исследовании живой лингвистической реальности, каковой являются тексты**. Применение АСА к Корпусу украинского языка дает возможность исследователям украинского языка установить синтаксическую и семантическую емкость такой единицы, как словосочетание, а в прикладном плане – разработанный автоматический синтаксический модуль анализа украинского текста – это механизм, с помощью которого становится реальностью составление по крайней мере двух словарей: частотного словаря словосочетаний и частотного словаря сочетаний простых предложений в сложном.

С помощью АГАТ-синтаксиса компьютер «поднимается» еще на одну ступеньку в процессе «понимания» содержания текста, приближаясь к разрешению конечной задачи АОТ – построению его семантического представления. Если АГАТ-морфология в терминах лексико-грамматических классов слов обеспечивает «понимание» компьютером денотативной информации, представленной в тексте, то АГАТ-синтаксис открывает путь к релятивной информации, т.е. к пониманию семантико-синтаксической структуры предложения. Синтаксические связи не существуют без семантических. И если не понятна синтаксическая структура предложения, не понятен и его смысл.

Задача **АГАТ-морфемики** заключается в возможности осуществлять лингвистические исследования морфемной и словообразовательной структуры, а именно: 1) составлять алфавитно-частотные словари всех типов морфов на базе текстов разных стилей и жанров; 2) объединять алломорфы в морфему; 3) устанавливать системные и функциональные характеристики морфем; 4) автоматически конструировать морфемно-словообразовательные гнезда. Информационной основой морфного сегментатора АГАТ-морфемики являются две базы данных – 170 тыс. слов и 3,5 млн. словоформ, в которых каждое слово и его словоформа представлены в виде морфной модели с информацией о типах морфов, их структурные отношения с другими морфами. В базовом словаре омонимичным корням (около 3100 единиц) и корневым алломорфам приписываются индексы из списка омонимических корней, а корневому алломорфу (около 2900) – инвариантная форма. Формализация морфных структур слов через описание их в терминах программных процедур морфной базы данных позволяет создавать на основе этой базы данных автоматизированную систему анализа, способную выполнять ряд таких **прикладных задач:** 1) группировать лексику по общеаффиксальным классам; 2) классифицировать лексику по морфным моделям; 3) создавать корневые и аффиксальные словари с учетом омонимии и алломорфии. Такое формализованное описание морфной структуры предусматривает моделирование структурных отношений морфов в двух планах организации слова как языкового знака: в плане выражения и плане содержания.

Работа морфемного модуля АГАТ-морфемики осуществляется самостоятельно и в связи со словообразовательным модулем с целью автоматического построения словаря морфемно-словообразовательных гнезд. Для этого: 1) группируется лексика в общекорневые выборки по процедуре идентификации корневых морфов, выделенных в словах морфемной базы; 2) на основе выборки всех общекорневых слов строятся словообразовательные гнезда как статьи электронного словообразовательного словаря на основе теоретических принципов словообразовательной производности. Формализованное описание морфных структур, предложенное при составлении электронного морфемного словаря, позволяет использовать морф как инструмент в проведении автоматичного морфемного анализа других лексикографических систем.

Компьютерная грамматика вообще и АГАТ-грамматика в частности, является динамическим механизмом, который является системой правил оперирования грамматическими значениями с акцентом на формы их выражения, с помощью которых компьютеру открывается доступ к денотативной и релятивной информации. **Компьютерные словарь и грамматика** – два тесно связанных и согласованных компонента структуры языка. Их согласованность определяется общностью основных функций и сохранением в компьютерной памяти как языковых единиц, готовых к употреблению, так и грамматических правил, по которым в соответствии с заданием автоматически осуществляется анализ текста. Результаты работы АГАТ-грамматики выложены для общего пользования на портале www.mova.info.

ЛИТЕРАТУРА

1. *Гончаренко, В. В.* Автоматическая переработка текста / В. В. Гончаренко [и др.]. – Кишинев : Штиинца, 1978. – 93 с.
2. *Зубов, А. В.* Основы искусственного интеллекта для лингвистов / А. В. Зубов, И. И. Зубова. – М. : Логос, 2007. – 319 с.
3. *Зубов, А. В.* Информационные технологии в лингвистике / А. В. Зубов, И. И. Зубова. – М. : Academia, 2004. – 205 с.
4. *Дарчук, Н. П.* Комп'ютерне анотування українського тексту: результати і перспективи / Н. П. Дарчук. – Київ : Освіта України, 2013. – 543 с.