Анжелла Лаврененко, Шен Хасан (Минск — Турция)

РАЗРАБОТКА БАЗЫ ДАННЫХ ПО СИНТАКСИСУ С ПОМОЩЬЮ EXCEL

Синтаксические исследования любого языка могут быть качественными, достоверными только тогда, когда основываются на большом лингвистическом материале, в идеале представляющем собой сплошную выборку языковых единиц по определенным параметрам из некоторого текста (текстов). В настоящее время возможность быстро получить требуемую выборку предоставляют корпуса текстов разных языков, в частности Национальный корпус русского языка. В составе НКРЯ имеется синтаксический подкорпус, из которого благодаря снятой грамматической омонимии можно получить некоторый языковой материал для синтаксических исследований. Однако проблема состоит в том, что поиск языковых единиц (предложений) во всем НКРЯ организован по слову, т. е. в синтаксическом подкорпусе будет доступна только информация о синтаксических свойствах слова в каждой из его словоформ. Предложение как языковая единица в данном корпусе не рассматривается.

Основной единицей синтаксического уровня языка является предложение. Исследователь нуждается в наиболее полном описании предложений по разным признакам с возможностью быстро извлечь из текста их все по заданным характеристикам. Преподаватель также должен иметь некоторую базу для подготовки к уроку. Таким образом, целью создания предлагаемой нами синтаксической базы данных стало облегчить синтаксистам подбор языкового материала в виде предложений с приписанными им характеристиками.

Для отработки технологии создания синтаксической базы данных был взят текст повести М.А. Булгакова «Собачье сердце», причем на первом этапе мы решили ограничиться простым предложением. Этот текст выбран не случайно: в нем представлены разнообразные речевые ситуации (авторское описание, диалог, монолог, внутренний монолог, полилог, дневниковые записи, телефонный разговор, наблюдаемый только с одной стороны). Мы предположили, что такое разнообразие речевых ситуаций даст также и разнообразие типов простых предложений.

Выборка простых предложений из текста и их описание производилось в полуавтоматическом режиме. Разбиение текста на предложения было сделано по формальным показателям (знаки препинания в конце фразы, следующая фраза начинается с заглавной буквы). Сложные предложения были исключены вручную. Таким образом, из текста повести было получено 1514 простых предложений.

В качестве инструмента для создания базы данных решено было использовать Excel, а не Access, потому что работа в Excel входит в программу по курсу «Информационные технологии», а для работы в Access нужны некоторые специальные навыки. Преимущество Excel в простоте

использования сортировки, в возможности неограниченного расширения параметров описания, их добавления по мере необходимости. Преимущество Access было бы в том, что сортировку можно производить по нескольким параметрам одновременно. Однако разработанные нами теги дают возможность быстро провести сортировку и при необходимости сохранить ее результаты на новом листе в том же файле, что удобно для описательной работы исследователя, для учителей при подготовке к уроку по определенной синтаксической теме, для составителей учебников и пособий.

Для характеристики предложений в нашей базе использовались традиционные признаки классификации, вынесенные в столбцы таблицы. Был разработан тег-сет в следующем виде:

Столбец B:
д — двусоставное
оо — односоставное определенно-личное
он — односоставное неопределенно-личное
об — односоставное безличное
оз — односоставное назывное
ов — вокативное
нч — нечленимое
Столбец C:
н — нераспространенное
р — распространенное
Столбец D:
н — неполное
п — полное
Столбец E:
в — восклицательное
н — невосклицательное
Столбец F:
п — повествовательное
в — вопросительное
б — побудительное
Столбец G:
од — осложнено однородными членами
об — осложнено обособленными членами
о — осложнено обращением
вв — осложнено вводной конструкцией
вс — осложнено вставной конструкцией
н — неосложнённое
м — имеет в составе междометие
п — имеет в составе повторы слов

Некоторые упрощения в классификации связаны с тем, что база данных создавалась магистрантом из Турции. Так, например, не выделялись оптативные предложения, поскольку морфологического выражения категория оптативности в русском языке не имеет. Инфинитивные предложения включены в состав безличных, но при необходимости могут быть выделены в отдельную группу. Кроме того, в типы осложнения мы добавили повтор слов (например, *Пропал я, пропал!, Вот он всё ближе, ближе.*) и междометия (сами по себе они представляют нечленимое

предложение, но их присоединение к простому предложению не делает языковую единицу полипредикативной; например: *О, бескорыстная личность. Эх, люди, люди!..*).

Каждое предложение нашей мини-базы данных было аннотировано соответствующими тегами. Восклицательные и вопросительные предложения аннотировались в автоматическом режиме, остальные параметры — вручную.

Особую сложность, как и предполагалось, вызвало аннотирование неполных предложений, и эта работа еще не завершена полностью, так как требует сверки с текстом. Возможно, в дальнейшем неполные предложения следует представлять в контексте. Но объем этого контекста далеко не всегда ограничивается одним-двумя предыдущими предложениями. В некоторых наших примерах объем необходимого контекста доходит до нескольких абзацев (например, неполные предложения с опущенным подлежащим в дневниках доктора Борменталя: *Фотографирован утром. Отчётливо лает «Абыр», повторяя это слово громко и как бы радостно. В 3 часа дня (крупными буквами) засмеялся, вызвав обморок горничной Зины. Вечером произнёс 8 раз подряд слово «Абыр-валг», «Абыр».* «Ближайшее» подлежащее находится в трех абзацах впереди по тексту). Встречаются случаи, когда восстановление опущенных членов предложения возможно не из предыдущего, а из последующего контекста: — *Чего ты? Ну, чего лаешься? — умильно щурил глаза пёс*.

Для целого ряда предложений затруднительно определить структуру даже в связке с контекстом, например: *Он засуетился, позвонил и вошедшей Зине приказал срочно подавать обед. — Обед! Обед! Обед*! Из ближайшего контекста следует, что *Обед!* — предложение двусоставное, неполное, распространённое. Но его можно рассматривать и как односоставное назывное (в таком случае — полное, нераспространённое). Характеристика по цели высказывания также затруднительна: побудительное или повествовательное.

Встречались случаи, когда из контекста структура неполного предложения не определяется:

Зина прибежала бледная.

— Зина, там в приемной... Она в приемной?

— В приемной, — покорно ответил Шариков, — зеленая, как купорос.

— Зеленая книжка...

— Ну, сейчас палить, отчаянно воскликнул Шариков, — она казенная, из библиотеки!

— Переписка — называется... как его... Энгельса с этим чертом... В печку ее!

Вероятнее всего, данное предложение следует рассматривать как односоставное неполное (или эллиптическое), но тип остается неясным: определенно-личное или безличное, инфинитивное.

Встретился случай, когда для определения структуры предложения понадобилось знание иного текста — стихотворения

А.С.Пушкина: *«От Севильи до Гренады...»* (пел профессор Преображенский). Однако у Пушкина эта строка входит в бессоюзное сложное предложение. Можно ли считать его простым в составе текста М. Булгакова?

В повести имеется такой фрагмент, как дневник, и в нем использованы предложения — записи дат: *24 декабря.* Его структуру определить однозначно также затруднительно: либо односоставное назывное, либо двусоставное неполное. Вопрос о распространенности также остается открытым: если это ответ на вопрос «когда?», то предложение будет распространенным.

Особую сложность представляют собой структуры, для которых невозможно определить даже количество предикативных центров, например: *Ни лампы в шелку, ни тепла, ни сытости*. Согласно «Русской грамматике», такие предложения следует относить к односоставным генитивным; следовательно, приведенный пример представляет собой бессоюзное сложное предложение с тремя предикативными центрами. Если же трактовать данное предложение как неполное (эллиптическое), с опущенным сказуемым «нет», то предложение оказывается простым осложненным.

Разрабатывая нашу синтаксическую базу данных, мы не ставили перед собой задачу дать ответы на все спорные вопросы. Поэтому многие единицы помечены в нашем корпусе знаком «?». Это дает возможность потенциальным исследователям быстро их вычленить из текста повести.

Разработанная нами база данных будет постепенно пополняться другими текстами с использованием этого же тег-сета. Так, силами студентов уже создается база данных по «Севастопольским рассказам» Л.Н.Толстого и «Повестям Белкина» А.С.Пушкина. Подобную синтаксическую базу данных можно разработать и для сложных предложений, включив туда иные параметры описания.