ИНТЕГРИРОВАННАЯ ИНФОРМАЦИОННАЯ СРЕДА STARLING И ЕЁ ИСПОЛЬЗОВАНИЕ В СФЕРЕ КОРПУСНОЙ ЛИНГВИСТИКИ

STARLING INTEGRATED INFORMATION ENVIRONMENT AND ITS USE FOR CORPUS RESEARCH

С. А. Крылов *(krylov-58@mail.ru),*

С. А. Старостин

Задачи корпусной лингвистики, решаемые системой STARLING: (1) преобразование письменного текста в многоуровневую текстовую базу данных (ТБД); (2) её разметка (автоматическая и полуавтоматическая) (с интерредактированием); (3) создание и корректировка словарных БД (с опорой на внешние источники).

0. STARLing и корпусная лингвистика

В докладе рассматриваются способы использования интегрированной информационной среды STARLing для решения некоторых важных (более того, на наш взгляд, основных) задач, стоящих перед корпусной лингвистикой. Имеется в виду создание текстовых баз данных (БД) и дальнейшая работа с ними, включающая создание вторичных – то есть, по сути дела, словарных баз данных.

Процедуры переработки текста, применяемые в корпусных исследованиях, образуют своеобразную технологическую цепочку. Ниже очерчены её важнейшие звенья.

1. Преобразование текста в стандартно организованную базу данных

Первый этап состоит в преобразовании письменного текста в текстовую базу данных (ТБД). Для решения этой задачи служит процедура т.н. «конвертирования» (CONVERT), а конкретнее, тот подраздел конвертирования, который называется «конвертирование текстового файла в файл БД».

Примечание. Термин «письменный текст» (далее - ПТ) понимается широко; в частности, он охватывает любую буквенную или иероглифическую фиксацию текста, первоначально произнесённого в устной форме, причём независимо от того, используется ли для записи общее (орфографическое) или специальное письмо (транскрипция, транслитерация).

Как известно, любой письменный текст характеризуется той или иной графической структурой. Разумно выделять несколько видов таких единиц. Так, типичными единицами членения ПТ являются: главы, параграфы, абзацы, стихотворные строфы, стихотворные строки, пунктуационные фразы, пунктуационные клаузы, пунктуационные синтагмы, графические словоформы.

Для каждого из этих видов членения ПТ используются особые виды разграничителей: символы нумерации глав и параграфов, абзацные отступы, сентенциальные делимитаторы (точка, восклицательный и вопросительный знак), клаузальные делимитаторы (точка с запятой, тире, двоеточие), синтагменные делимитаторы (запятые), словесные делимитаторы (пробелы). Таким образом, один ПТ имеет (обычно) несколько графических структур разных уровней, и эти структуры, как правило, друг с другом не совпадают. Поэтому одному тексту можно, вообще говоря, поставить в соответствие не одну ТБД, а несколько разных: «членение на главы», «членение на абзацы», «членение на пунктуационные фразы» и т.п., вплоть до «членения на графические слова». Есть две стратегии построения ТБД: одноуровневая и многоуровневая. При одноуровневой стратегии в ТБД содержится в явном виде информация лишь о каком-то одном уровне членения текста. При многоуровневой стратегии в ТБД содержится информация одновременно о нескольких (в предельном случае – даже обо всех) уровнях членения ПТ. При создании ТБД целесообразна стратегия последовательного расчленения текста – сначала на более крупные единицы членения, потом на единицы более мелкие и т. д.

Если набор делимитаторов однозначно задан в тексте, то соответствующий ему вид разметки осуществляется в автоматическом режиме.

При осуществлении надлежащих процедур, расчленяющих ПТ на более тонкие и нетривиальные единицы можно получить такую ТБД, в которой бы содержалась информация и о соответствующих видах членения. Так как в последнем случае набор соответствующих делимитаторов в ПТ отсутствует, то и разметка ПТ оказывается возможной лишь с опорой на интуицию разметчика. Следует, однако, подчеркнуть, что техника представления текста в виде ТБД позволяет облегчить труд разметчика, дав ему возможность вместо чисто ручной процедуры разметки осуществить «полуручную» (= «полуавтоматическую») разметку.

Так, между графическим словом и пунктуационной синтагмой можно выделить особый уровень, который

(за неимением лучшего термина) назовём уровнем «синтаксических молекул» (используя термин Ш. Балли). Синтаксической молекулой при таком подходе будет считаться сочетание одного знаменательного слова с одним или несколькими служебными словами, линейно примыкающими к нему (обычно такой тесный контакт сопровождается и некоторой степенью фонетической слитности, т.н. «клитизации», и некоторой степенью семантической слитности – формирование единого «члена предложения»). С точки зрения фонетики синтаксические молекулы обычно соответствуют единицам, которые принято называть фонетическими словами.

Если в распоряжении создателя ТБД есть словарь морфем исследуемого языка и сведения о внутренней структуре словоформ, то возникает принципиальная возможность представить текст как цепочку морфем.

Если в распоряжении исследователя есть достаточно полный фразеологический словарь исследуемого языка, то возникает принципиальная возможность фразеологической разметки ТБД, при которой единицей членения текста будет «идиома» (в смысле Ч. Хоккета) или «инвентарная единица» (в смысле В. Б. Касевича), то есть единица, хранимая в некотором относительно ограниченном инвентаре языковых единиц.

2.Автоматическая разметка ТБД

Автоматическая разметка ТБД возможна для разных типов информационной структуры ПТ. Она включает в себя, в частности, морфологический, синтаксический и семантический анализ языковых единиц.

Наибольшей полноты сегодня достигает автоматический морфологический анализ. Он осуществляется полностью автоматически специализированным модулем системы STARLing. В настоящее время STARLing содержит морфологические анализаторы двух языков – русского и английского. Объём сведений о морфологической структуре русских словоформ опирается на «Грамматический словарь русского языка» А. А. Зализняка. Объём сведений о морфологической структуре английского языка соответствует информации, содержащейся в англо-русском словаре В. К. Мюллера.

В отличие от некоторых других систем автоматического морфологического анализа текста, анализатор системы STARLing может функционировать в качестве компонента автоматизированного рабочего места (АРМ) лингвиста. Это проявляется, в частности, в том, что STARLing не только выводит «на экран» результат морфологического анализа словоформы (а также генерирует на экране словоизменительную парадигму соответствующей лексемы), но и содержит встроенную функцию, возвращающую результат морфологического анализа словоформы – функцию GRAMMAR. Благодаря такой функции можно выполнить, в частности, следующую задачу: имея на входе произвольную базу данных, соответствующую цепочке словоформ данного (русского или английского) языка, получить на выходе базу данных, содержащую её пословный морфологический разбор.

В системе STARLing имеется и модуль автоматического синтаксического анализа (АСАн), опирающийся на те сведения о синтаксических свойствах русских слов, которые содержатся в словаре А. А. Зализняка. Содержательно модуль АСАн базируется на принципах грамматики зависимостей (восходящей к идеям А. М. Пешковского и Л. Теньера); во многом он сближается с теорией формального синтаксиса русского языка, развитой И. А. Мельчуком в его работах по АСАн 1960-х и 1970-х гг.

Несмотря на то, что многие случаи синтаксической неоднозначности приводят к выдвижению нескольких альтернативных разборов, выбор между которыми возможен пока лишь вручную, остаётся принципиальная возможность постепенного подключения разнообразных «фильтров», позволяющих отсеивать семантически рассогласованные варианты разбора и тем самым минимизировать количество альтернативных синтаксических разборов.

Модуль лексико-семантического (в том числе лексико-стилистического и отчасти лексико-синтаксического) анализа опирается в настоящее время на базу данных о русской лексике на основе “Словаря русского языка” С. И. Ожегова). С помощью этого модуля можно обработать русский текст произвольной длины, снабдив его автоматической семантической аннотацией. Семантическая аннотация текста в настоящее время представляет собой цепочку семантических аннотаций ко всем словоформам, из которых этот текст составлен. В случае лексической омонимии и полисемии словоформ семантическая аннотация может быть «малой» и «большой». «Малая» аннотация содержит толкование «основного («прямого») значения данного слова; «большая» аннотация содержит толкования всех значений данного слова (как прямого, так и всех переносных словарных значений).

Несмотря на то, что семантическое аннотирование словоформ в автоматическом режиме, как и вообще любой автоматический анализ (при любой работе с корпусами в условиях «неснятой омонимии»), создаёт известную долю «информационного шума», этот путь семантического анализа текста представляется вполне допустимым путём извлечения информации из текста. В частности, система поиска информации, опирающаяся на такие аннотации, может успешно извлекать из текста информацию об объектах, входящих в некоторый класс, по запросу, содержащему наименование этого класса, при условии, если словарное толкование лексемы со значением разновидности чего-то содержит вхождение естественно-языковой лексемы, обозначающей понятие, родовое по отношению к понятию, обозначаемому данной лексемой.

3. Полуавтоматическая разметка ТБД

Разметка ТБД в полуавтоматическом режиме предполагает применение интерредактирования. Этот вид металингвистической деятельности оказывается весьма существенным при обработке крупных текстовых массивов.

При допущении интерредактирования оказывается возможным задание границ фонетических слов (в частности, “тактов” //“акцентных слов”) в ПТ на основе некоторого изначально заданного представления о составе множества клитик и клитикообразных слов (на основе такой разметки можно производить их классификацию, составление частотных словарей тактов и т. п.). Так, для русского языка оказывается возможным выделить по меньшей мере следующие 6 видов скрытых делимитаторов («пограничных сигналов», «стыков»), выделяемых внутри «макротакта» (слитно произносимого отрезка речи, внутри которого нетипична или невозможна пауза): граница проклитики и её правого соседа; граница энклитики и её левого соседа; граница проклитикоида и его правого соседа; граница энклитикоида и его левого соседа; граница между компонентами квази-композита с неустойчивой позицией просодического центра; граница между компонентами клишированного (относительно устойчивого) оборота. При расстановке символов соответствующих делимитаторов происходит чередование процедур ручной и автоматической разметки. Так, границу между первообразным предлогом и последующим словом можно в большинстве случаев автоматически разметить как «проклитическую», однако разметка границ при частице «*ведь*» (при анализе ПТ без интонационных маркеров) не может быть поручена компьютеру, так как требует обращения к интонационной структуре синтагмы.

К полуавтоматическим процедурам относится снятие морфологической омонимии в корпусе. Так, если синтаксический контекст позволяет разрешить морфологическую омонимию, то возможен вывод на экран всех записей, удовлетворяющих тому или иному условию, задаваемому в терминах синтаксических фильтров. После такой фильтрации задача снятия морфологической омонимии упрощается (и т.о. убыстряется) в несколько раз, как показывает практика морфологической разметки русскоязычных корпусов.

Опять же именно к полуавтоматическим процедурам можно отнести расстановку однозначных помет о синтаксических связях в тексте. Удобный технический приём работы с синтаксическими связями – это вывод на экран контактных или близко дистанцированных сочетаний словоформ, обладающих определёнными заранее заданными морфологическими характеристиками. Фильтрация такого типа позволяет разметчику достаточно быстро справляться со зрительной обработкой вхождений элементарных типовых сочетаний грамматических форм, представленных в ТБД большого объёма.

К полуавтоматическим процедурам можно отнести многие случаи разрешения лексико-семантической омонимии и полисемии. Такая ситуация имеет место в тех случаях, когда удаётся сформулировать правила выбора значения с опорой на тот или иной тип контекста (грамматического – с опорой на результаты морфологического анализа; синтаксического – с опорой на результаты синтаксического анализа; лексического – с опорой на лексико-семантические аннотации при словоформах) и т. п.

4. Реструктурирование ТБД

После того, как ТБД создана, она может быть подвергнута реструктурированию и обогащению. Для этого используются специальные операции, функции и процедуры.

С точки зрения своего генезиса STARLing как система управления базами данных (СУБД) представляет собой продукт развития той же программистской традиции, к которой принадлежат, напр., СУБД серий Dbase, FoxBase, FoxPro и т. п. В настоящее время большая часть системы STARLing написана на языке Clipper в сочетании с языком C. Поэтому многие стандартные функции и операции системы STARLing восходят к соответствующим элементам традиции Dbase.

Наиболее существенный вид реструктурирования ТБД – это изменение её структуры, осуществляемое с помощью процедуры «Изменение структуры» (Modify structure). Эта процедура позволяет добавлять или удалять те логические поля, в которых записывается информация; изменять физический порядок полей, а также (при необходимости, что бывает редко) изменять логический тип того или иного поля.

Обогащение ТБД может осуществляться с помощью разнообразных процедур, операций и функций.

В отличие от текстовых редакторов, где единственная разновидность замены, не требующая составления специальных макросов, – это стандартная замена подцепочек текста, в среде STARLing замена осуществляется несколькими типами стандартных операций, важнейшими из которых являются «текстовая замена» (REFORM) и «логическая замена» (REPLACE).

Процедура «логической замены» (REPLACE) позволяет заменить содержимое любого поля выражением, которое может быть простым (представлять собой константу или простую переменную) или сложным – т. е. представлять собой значение некоторой простой или сложной логической функции от некоторых аргументов.

В качестве констант выступают числа, даты, истинностные значения и символьные цепочки. В качестве простых переменных выступают многие виды сущностей, но для пользователя-филолога важнейший вид простых переменных – это переменные со значением логических полей, а также переменная RECNO().

Логические поля в ТБД автор ТБД создаёт по своему усмотрению. Например, можно создать отдельные поля для таких видов информации, как «орфографическая запись словоформы»; «транскрипция словоформы»; «заглавная форма лексемы»; «словоизменительная характеристика (грамматическая форма) словоформы»; «падеж словоформы»; «грамматическое число словоформы»; «грамматическая одушевленность словоформы»; «репрезентация словоформы»; «наклонение словоформы»; «время словоформы»; «залог словоформы»; «грамматический вид словоформы»; «грамматическое лицо словоформы»; «предикационный статус словоформы»; «часть речи лексемы»; «грамматический подкласс лексемы», «словоизменительный подкласс

лексемы»; «акцентный тип лексемы»; «перевод лексемы на чужой (например, английский по отношению к русскому) язык (например, согласно русско-английскому указателю к словарю В. К. Мюллера)»; «толкование данной лексемы по словарю С. И. Ожегова»; «стилистическая характеристика данной лексемы по словарю С. И. Ожегова», «синтаксическая роль данной словоформы относительно её синтаксического хозяина»; «синтаксическая роль данной словоформы относительно её N-го (1-го, 2-го, 3-го) синтаксического слуги» «линейная позиция данной словоформы относительно её синтаксического хозяина»; «линейная позиция данной словоформы относительно её N-го (1-го, 2-го, 3-го) синтаксического слуги»; «линейная позиция данной словоформы относительно антецедента или ближайшей слева словоформы, кореферентной данной словоформе», «линейная позиция данной словоформы относительно N-го (1-го, 2-го, 3-го) консеквента или ближайшей справа словоформы, кореферентной данной словоформе», «семантическая помета, объясняющая данную словоформу с помощью её перифразирования средствами литературного языка (для диалектизмов)» и т. п.

Переменная RECNO() обозначает номер данной записи в данной БД.

Среди логических функций, владение которыми полезно для работы филолога, особое значение имеют такие:

«Подцепочка» (SUBSTR) по отношению к некоторой объемлющей цепочке, т.е. её часть от такой-то позиции длиной во столько-то символов.

«Позиция» (AT) некоторой подцепочки в составе объемлющей цепочки.

«Длина» (LEN) цепочки.

«Количество вхождений» (HOWMANY) некоторой подцепочки в состав объемлющей цепочки.

«Зеркальное оборачивание» (REVERSE) цепочки; ценность этой функции становится очевидной для лингвистов, когда-либо работавших с обратными словарями, а особенно для тех, кто имеет собственный опыт составления обратных словарей.

«Символьная запись» (STR) некоторого числа.

«Числовое значение» (VAL) некоторой цепочки.

При построении сложных логических выражений из простых используются стандартные логические операторы «не» (.NOT. или !); «и» (.AND.); «или» (.OR.). К ним примыкают такие операторы, как «если» (IF) и некоторые другие.

Особую важность при работе с БД различных типов имеет функция «содержание записи» (в некотором поле или во всех полях) (RECORD). С помощью этой функции можно задавать выражения, производные от содержания полей не только в данной записи, но и во всех записях, содержащихся в БД. Тем самым образуется возможность формального вычисления и фиксации (в виде значений эксплицитно задаваемых функций) любых видов информации, прямо или косвенно выводимых из содержания любых записей в любых полях.

Для ТБД наличие функции «запись» означает возможность гибкой фиксации (и потому дальнейшего эффективного использования) любых синтагматических отношений между языковыми единицами, вхождения которых образуют структуру текста и любых его компонентов.

Для словарных БД наличие функции «запись» означает возможность гибкой фиксации (и потому дальнейшего эффективного использования) в работе любых видов информации, задаваемых не в данной записи, а в предшествующей, или предшествующей на N записей назад, или последующей (напр., последующей через N записей вперёд) и т. п.

5. Поиск информации в ТБД

ТБД как продукт обработки текста представляет эвристическую ценность не только как промежуточный продукт, на основе которого строятся вторичные БД, но и как удобный инструмент поиска информации в тексте.

Для удобства поиска информации в системе STARLing используется несколько разных поисковых операций наряду с множеством специальных функций.

Удобным инструментом просмотра БД являются индексы. Каждый индекс представляет собой информацию о том, в каком порядке будут выводиться (на экран, на принтер или в специальный файл) записи текущей БД. В повседневной работе филолога удобно пользоваться такими типами индексов, как прямой и обратный алфавитный, частотный по убыванию, частотный по возрастанию, тематический (систематический), алфавитно-гнездовой, хронологический, географический и т. п. В среде STARLing индексирование осуществляется с помощью операции «индексирования» (INDEX).

Простейший вид поиска (операция SEARCH) устроен так же, как поиск подцепочек в текстовых редакторах. Однако STARLing позволяет добавить к этому виду поиска также «поиск по индексу» (операция SEEK) и «логический поиск» (операция LOCATE).

«Поиск по индексу» возможен лишь при наличии «индекса».

«Поиск по индексу» позволяет быстро находить местонахождение записей, виртуально упорядоченных в соответствии с некоторой линейной иерархией (типичный пример такой иерархии: прямой алфавитный порядок следования записей в ТБД).

«Логический поиск» позволяет находить записи, удовлетворяющие определённому логическому условию. Типичные примеры логического поиска – поиск информации о вхождениях определённых лексем и определённых грамматических форм.

6. Создание вторичных БД и их использование

На основе ТБД могут создаваться разнообразные «вторичные» БД. Для этого в среде STARLing используются такие процедуры, как «создание новой БД» (Create DBF File), «изменение структуры» БД, «сложение» БД (COMPOSE), «разложение» БД (DECOMPOSE), «подстановка» (SUBSTITUTE), а также такие операции, как «сортировка» (SORT), «текстовая замена» и «обновление» (UPDATE).

При металингвистическом индексировании БД полезна опора на электронные версии традиционных “бумажных” словарей (двуязычных, академических толковых, синонимических, фразеологических, орфоэпических, ортологических, синтаксических, а также энциклопедических).