ОТ СИСТЕМЫ СКАЗКА К СИСТЕМЕ СКАЗКА-2
FROM COMPUTER SYSTEM SKAZKA TO SKAZKA-2

А.В. Рафаева anna_raf@rambler.ru

НИВЦ МГУ

Рассматривается система СКАЗКА, построенная на основе указателя сказочных типов Аарне-Томпсона, обсуждаются ее структура, использование при анализе сказочных сюжетов и те изменения, которые необходимы при переходе к анализу фольклорных текстов (система СКАЗКА-2).

Система СКАЗКА разрабатывается автором с 1996 г. прежде всего как справочное и исследовательское средство, которое можно использовать для анализа сюжетов волшебных сказок. Система реализована в СУБД STARLING и может выполняться как под управлением DOS, так и в Windows. Основой системы служит фрагмент указателя сказочных типов Аарне-Томпсона ([1], далее АТ), содержащий сокращенные описания сюжетов сказок на английском языке и некоторую справочную информацию.

Основной целью, которая преследовалась при создании системы СКАЗКА, являлась автоматизация ряда трудоемких операций по нахождению сказочных сюжетов, отвечающих некоторым условиям, например, таким, в которых действуют определенные персонажи или встречается какой-то мотив из числа тех, выделение которых формализовано в системе. Кроме того, в ряде случаев система использовалась для изучения сказочной картины мира (например, системы родственных отношений в волшебной сказке). Кроме того, естественно, система могла в дальнейшем использоваться для проведения исследований, предусмотреть которые заранее не представлялось возможным. Все это требовало разработки гибкого настраиваемого средства, позволяющего легко добавлять новые возможности и данные по мере необходимости. В частности, требовалось обрабатывать некоторые поля БД (например, поля, содержащие список мотивов по указателю Томпсона, появляющихся или могущих появиться в рассматриваемом сюжете) как текст и как единицы описания более высокого уровня в зависимости от требований конкретного исследования, необходимыми представлялось наличие средств, позволяющих проводить поиск записей, удовлетворяющих некоторым условиям (в том числе и списка условий, представленных в виде отдельной БД), выделение таких записей для дальнейшего изучения и сбор статистической информации по вводимому условию. Кроме того, хотелось также иметь процедуру, позволяющую частично автоматизировать проверку и первичную обработку вновь вводимой информации. Все эти возможности в системе реализованы (частично они предоставляются средой STARLING, частично реализованы в виде отдельных программных модулей, написанных автором)

Ядром системы является текст указателя АТ, представленный полностью в виде БД с текстовыми полями. Основная часть исследований и все эксперименты над указателем АТ как над текстом на ЕЯ проводились именно над этим ядром, к которому, в зависимости от целей конкретного исследования, могу т подключаться некоторые вспомогательные БД, постоянные или временные. К их числу относятся:

- Список фольклорных мотивов по указателю С. Томпсона с указанием сказочных типов, в которых встречаются данные мотивы (составляется автоматически по материалам АТ);

- Временная БД, в которые вносятся конкордансы любого слова, заданного пользователем (составляется с помощью отдельно запускаемой программы);

- Ряд БД, включающих дополнительные сведения о сказочных типах, внесенные исследователем. Каждая такая БД разрабатывается отдельно для конкретных исследовательских задач и связывается с основной БД (ядром) собственными правилами.

- БД, содержащая список ключевых слов (КС) и их соответствий, т.е. соответствий вида *ключевое слово – явление (список явлений), о наличии которого может сигнализировать появление данного КС.* Некоторые соответствия могут выполняться не всегда, а при выполнении какого-то предварительного условия. Например, КС *transform* в тексте указателя сигнализирует о наличии мотивов *заколдование* или *расколдование,* однако в случаях, когда ключевое слово *enchant* предшествует данному, можно с вероятностью, близкой к единице, считать, что в тексте представлен именно мотив *расколдование*. В настоящее время описание этих условий не формализовано и представляется для пользователя, составляющего запрос к БД, а не для компьютерной обработки.

Большая часть компьютерных экспериментов производится с помощью запросов к основной и вспомогательным БД на языке запросов SQL – либо непосредственно в виде запросов, либо в виде небольших программных модулей, исполняемых средой STARLING. Так, с помощью подобных модулей реализованы функции составления БД конкордансов заданного слова, выделения мотивов по указателю Томпсона в

отдельную БД, связанную с основной и т.д. Однако для некоторых задач, прежде всего, для задач, требующих большого объема работы с текстовыми данными, возможностей, предоставляемых встроенным интерпретатором, недостаточно. К таким задачам относится, к примеру, составление частотного словаря всех словоформ, встретившихся в тексте указателя. Для решения таких задач используются программные модули, реализованные на языке C++.

Наиболее удобным и плодотворным для анализа как сюжета, так и картины мира волшебной сказки, оказался метод поиска и выделения ключевых слов. В качестве ключевых слов выступает последовательность символов (не обязательно именно слово, это может быть часть основы слова или словосочетание), которое с высокой степенью вероятности сигнализирует о появлении в тексте некоторой единицы классификации. Таким образом, использование ключевых слов, по сути дела, не дает нам полной картины строения текста, но позволяет достаточно быстро найти и выделить для исследования именно те тексты, в которых исследуемое явление представлено (поиск с использованием только части ключевых слов, дающих точные результаты) или может быть представлено (поиск с использованием всех ключевых слов).

К сожалению, в настоящее время автоматическое нахождение всех или только некоторых ключевых слов не представляется возможным; в то же время поиск необходимых ключевых слов только вручную даст нам заведомо неполные списки ключевых слов, хотя и отличающихся высокой степенью точности. Одним из средств, позволяющих частично автоматизировать работу по выделению КС и проверке его роли в волшебной сказке, является составление отдельной БД его конкордансов с последующим анализом этой БД и ручным определением роли данного КС в сюжете. Другими средствами, частично автоматизирующими выделение КС, является использование частоты встречаемости слов (всех словоформ данного слова) в тексте указателя и попеременная работа с англо-русским и русско-английским словарем, входящими в состав среды STARLING. К сожалению, частотный анализ встречаемости слов в тексте не позволяет выделить так называемые «нули», т.е. те ключевые слова и понятия, которые значимы для фольклорной картины мира, но редко встречаются в текстах данного жанра. К примеру, для волшебных сказок такими «нулями» будут практически все термины непрямого родства – в отличие от терминов прямого родства и свойства (отец, мать, дети, брат и сестра, муж, жена, жених, невеста).

Кроме того, поиск КС затрудняется еще и тем, что в качестве ключевых часто выступают не лексемы, а мифемы, т.е. понятия, которые можно описать с помощью пучка дифференциальных признаков на основе ряда значимых в сказке семантических оппозиций. Например, *царевна* может быть описана с помощью системы оппозиций *молодой*/*взрослый*, *состоящий*/*не состоящий в браке* и *высокий/низкий* (с социальной точки зрения) как молодая девушка, занимающее высокое социальное положение. В ряде сказок с помощью тех же оппозиций будет описана (и, следовательно, будет выполнять ту же роль в сюжете) купеческая дочь, поповна и т.п. В то же время младший из трех царских сыновей (часто Иван-царевич) будет играть низкую роль (по положению в семье) по сравнению со своими братьями, что для волшебной сказки более является более значимым, чем его очевидно высокое социальное положение. Таким образом, выделение мифем (или даже их части) невозможно без полного анализа сюжета сказки и определения ролей сказочных персонажей относительно друг друга.

Опыт работы с системой СКАЗКА позволил сделать следующие выводы:

1. Определены свойства указателей, которые делают их удобными для автоматической обработки (прежде всего это единообразие описания, достаточно полные описания сюжетов. В частности, указатель [2] неудобен для автоматической обработки именно из-за коротких описаний сказочных сюжетов. Кроме того, как оказалось, более удобны для автоматического анализа те фольклорные указатели, в которых описан не один, а несколько жанров. При этом принятая автором система классификации не играет большой роли для автоматической обработки.
2. Выделение текстов, включающих определенные КС, позволяет найти тексты, обладающие заданными свойствами.
3. Найден ряд методов, позволяющих частично автоматизировать поиск и выделение КС, относящихся к заданной области.
4. Для работы с указателями использование статистических методов анализа малорезультативно.

В настоящее время нами разрабатывается система СКАЗКА-2, основой которой является корпус текстов волшебных сказок, начиная с собрания сказок А.Н. Афанасьева, а указатели (прежде всего [2]) будут служить только в качестве вспомогательной справочной базы. Основными чертами, отличающими новую версию системы, являются следующие:

1. Опора на анализ фольклорных текстов, а не текстов из указателя;
2. Более сложная структура как баз данных, так и связей между отдельными таблицами БД;
3. Использование более сложных алгоритмов анализа текста, по всей вероятности с привлечением дополнительных словарей и других источников;

4. В дальнейшем предполагается распространение системы СКАЗКА-2, что делает вторую версию системы более требовательной к пользовательскому интерфейсу.

Эти отличия усложняют структуру системы. Поэтому предполагается также создание программной оболочки, которая позволит работать с текстами сказок как напрямую через СУБД STARLING, так и запуская дополнительные программы обработки текста.

.

Литература:

1. Thompson S. The Types of the Folktale: A Classification and Bibliography. Anti Aarne's Verzeichnis der Märchentypen. Third printing. Helsinki, 1973. FF Communications No. 184.

2. Сравнительный указатель сюжетов. Восточнославянская сказка / Сост. Л.Г. Бараг и др. Л., 1979. Доступно на сайте «Фольклор и постфольклор» http://ruthenia.ru/folklore/sus/index.htm