Формальная модель и база данных по русскому согласованию[1]

В. Д. Соловьев, А. М. Гусенков
Казанский госуниверситет, Институт проблем информатики АН РТ, Казань
solovyev@mi.ru

Аннотация. В докладе описывается формальная модель согласования. Модель строится на данных русского языка, но, предположительно, является универсальной. Модель позволяет единообразно трактовать различные конструкции согласования и дать объяснение сложным случаям согласования. В докладе также описывается компьютерная база данных согласования в русском. Она сопоставляется с аналогичной базой, реализованной в Суррее. Наша база данных ориентирована, с одной стороны, на предложенную модель, а с другой стороны, на описание согласования у Л. Иомдина.

Введение

Целями работы являются: 1) наметить контуры формальной модели согласования, приближающейся по степени строгости и методологии построения к естественнонаучным теориям, 2) описать компьютерную базу данных по русскому согласованию.
Основные требования, предъявляемые к модели.
а) Большое количество наблюдаемых явлений (в идеале все в заданной области) должно описываться исходя из небольшого числа базовых допущений.
б) Модель должна не только описывать, но и объяснять имеющиеся факты.
в) Модель должна обладать предсказательной силой. Применительно к лингвистике это означает следующее. Модель должна выявлять закономерности, которые могут быть сформулированы в виде универсалий (предсказание для тех языков, которые не были учтены непосредственно при построении модели).
Таким образом, модель ориентирована на теоретическое исследование структуры русского языка и человеческих языков вообще. Цель - выйти на уровень когнитивных механизмов language device.
Модель строится на основе данных русского языка. Рассматривается только согласование в роде и числе. В данном изложении не охватывается всё согласование в русском языке. Акцент сделан на ключевые моменты, позволяющие наиболее четко выявить особенности предлагаемой методологии. Изложение носит конспективный характер. Модель строится по шагам с последовательным усложнением.
Согласование является сложным явлением, описываемым во многих языках мира большим числом правил. Естественным является построение баз данных, которые позволяют

[1] Работа выполнена при поддержке РФФИ, грант N 02-07-90230

унифицировать описания и создают прочную основу для контрастивных и типологических исследований.
Хорошо известна база по согласованию [1], созданная в Университете Суррея (Великобритания) и содержащая, в том числе, данные по русскому языку. Русское согласование описано в ней с помощью нескольких десятков правил; описания даны на английском языке.
Наша база данных имеет описания и интерфейс на русском языке, что облегчит ее использование российскими лингвистами. Далее, описание согласования в английской базе данных является далеко не полным. Более детальное описание имеется в монографии Л. Л. Иомдина [2]. Создаваемая база данных ориентирована на эту монографию. Структура базы данных близка к разработанной в Англии с некоторыми уточнениями, вытекающими из работы Л. Л. Иомдина и описываемой ниже формальной модели. В дальнейшем планируется расширить базу данных на татарский язык.

1. Формальная модель согласования

1.1. Взаимодействие синтаксического и семантического уровней

Большинство авторов (Иомдин, Мельчук, Плунгян) трактуют согласование как поверхностное морфо-синтаксическое явление и стремятся четко разграничить семантическое и синтаксическое согласование. Однако под влиянием работ по искусственному интеллекту естественно рассмотреть и иную точку зрения – возможно, что в обработке предложений когнитивной системой человека семантический и синтаксический уровни очень тесно переплетаются.
Поэтому наша модель строится исходя из равноправного участия в согласовании синтаксической и семантической информации. Согласование подлежащего с глаголом и существительного с определением описывается в нашей модели единообразно. Для наглядности диаграмм, на них изображается только один вид согласования (начнем со сказуемого).

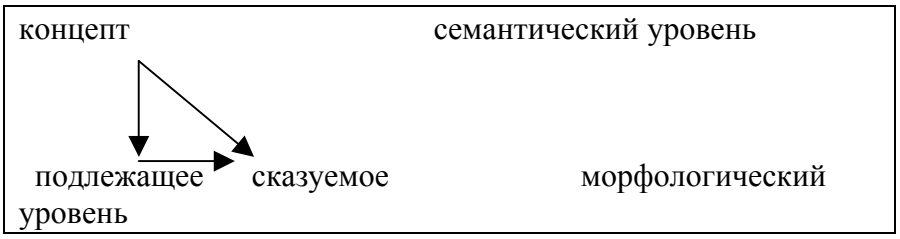

Диаграмма 1. Схема передачи информации при согласовании.

Основные постулаты модели.

1. Семантическая информация о числе/роде вербализуемого концепта всегда передается на морфологический уровень и соответствующему существительному, и глаголу.
2. Информация о числе/роде подлежащего передается в точку, в которой определяется форма глагола.
3. При определении формы глагола учитываются оба поступающих значения и выбирается одно из них.

Будем называть ‘подлежащее → сказуемое’ морфологическим каналом передачи информации и обозначать буквой m (как и значение, передаваемое по этому каналу). А ‘концепт → сказуемое’ назовем семантическим каналом и обозначим – s. С точкой, в которой определяется форма сказуемого, ассоциируем функцию выбора Ch(m,s) от соответствующих переменных. Ставится задача описать функцию Ch.

Если по обоим каналам поступает одно и то же значение, то проблем не возникает – это бесконфликтный случай, на который приходится большая часть всех согласований. На языке теории категорий из математической логики можно сказать, что диаграмма 1 коммутативна.
Если по обоим каналам поступает разная информация, то возникает конфликт, который и разрешается функций выбора. Рассмотрим два примера согласования в числе.
Если подлежащее *ножницы*, то m = мн. число и независимо от значения s (реального числа ножниц), сказуемое ставится во множественном числе, что на введенном метаязыке выражается, как Ch (m, s) = m.
Если подлежащее *белье*, то здесь, наоборот, m = ед. число, а семантика, скорее всего множественная (несколько "экземпляров" белья), т. е. s = мн. число. Сказуемое ставится в единственном числе, т.е. опять оказывается, что Ch (m, s) = m.
Таким образом, для любых значений m и s имеем:
(*) Ch (m, s) = m
Пока что описание не отличается по результату от традиционного, но в этой модели порции семантической и морфологической информация рассматриваются как равноправные в точке, в которой принимается решение о форме сказуемого.

1.2. Стековая модель: функция выбора от неопределенных аргументов

Что произойдет, если m не определено? В частности, это имеет место для параметра 'род' личных местоимений. В этом случае форма сказуемого определяется только на основе семантической информации. Обозначим это следующим образом:
(**) Ch (,s) = s
Аналогичным образом можно объяснить и согласование (в числе) со словами типа *пальто*. Мы считаем, что по морфологическому каналу информация о числе слова *пальто* не передается, т. е. что оно не имеет числа.
Обычно в грамматиках принимается альтернативная точка зрения, что неизменяемые слова имеют совпадающие формы ед. и мн. числа. Это, однако, никак не объясняет, как определяется форма сказуемого. Т. к. по форме слова *пальто* невозможно определить, сколько их, то все равно придется обращаться к семантической информации. В терминах предлагаемой модели этот вариант можно описать, считая, что по морфологическому каналу передаются одновременно два значения и ед. и мн. Но с точки зрения принятия решения о форме сказуемого, нет никакой разницы - не передается никакой информации или передаются сразу все возможные значения - это не уменьшает исходной неоднозначности, т. е. ровно ничего не дает.
Опишем теперь устройство для вычисления функции Ch. Представим себе стек, способный хранить только одно значение и очередь на попадание в него, как изображено на диаграмме 2. В начальный момент стек пуст.

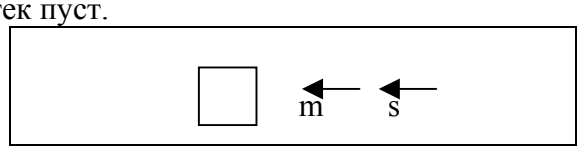

Диаграмма 2. Устройство для вычисления функции выбора

Ясно, что это схема вычисляет функцию Ch по формулам (*) и (**).
Что произойдет, если нет ни m, ни s? Такая ситуация наблюдается в безличных предложениях: *Пулей убило бойца.* Здесь нет не только морфологической информации о подлежащем m, но и семантической информации (кто убил бойца – Бог (мж.), провидение (ср.), карма (жн.)?).
Для описания этого класса случаев, увеличим размер стека до 2 элементов и положим, что стек изначально хранит 'дефолтное' значение 'един. число-ср. род'.

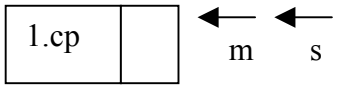

Диаграмма 3. ‘Дефолтное’ заполнение стека

Предварительный итог. На основе допущений (I-III), визуализированных на диаграмме 1, и с помощью механизма выбора, реализованного стековой схемой диаграммы 3, удается единообразно описать и объяснить группу различных (с точки зрения существующих грамматик) явлений.

1.3. Дальнейшее развитие модели

В русском языке имеется много случаев согласования, не укладывающихся в схему диаграммы 3 и требующих развития более сложной модели. Например, как объяснить генерацию морфологических форм и саму возможность варьирования в таких примерах как: *пришли 10 человек* и *пришло 10 человек*?

Мы отложим рассмотрение этих вопросов до другой статьи, констатировав здесь лишь тот факт, что реалистическая теория согласования должна безусловно учитывать как морфологическую так и семантическую информацию и описывать их нетривиальное взаимодействие при синтезе словоформ.

2. Структура базы данных

Структура нашей базы данных в основном совпадает с предложенной в [1]. Запись базы данных содержит следующую информацию: контролер, цель, категория контролера, категория цели, значения категории контролера, значение категории цели, синтаксическое отношение, условия согласования, тип согласования, условия выбора типа, пример. Здесь использована терминология принятая при описании согласования.

Для примера перечислим контролеры согласования для синтаксического отношения “подлежащее-сказуемое”. Это – именная группа, личное местоимение, инфинитив, подчиненное предложение, именная группа в косвенном падеже, количественная именная группа и три вида опущенния контролера.

Основные отличия от базы данных [1] состоят в следующем.

1) Наша база данных содержит описание большего числа редких случаев и исключений.

2) Различаются трактовки понятия “согласование”. Английские исследователи относят к согласованию и такое явление, как выбор формы анафорических местоимений. Это не бесспорно, и в нашу базу данных не включено.

3) В нашей базе данных введен дополнительный параметр – тип согласования – принимающий значения: ‘синтаксическое’ и ‘семантическое’, в соответствии со значением функции выбора Ch.

Реализация базы данных. Создано локальное сетевое приложение в среде СУБД Visual FoxPro, реализующее интерфейс пользователя с базой данных. В приложении реализована возможность выборки интересующих пользователя типов согласований. Пользователь имеет возможность выбора на экранной форме значений из справочных таблиц (в том числе и пустых значений). На основе этой выборки генерируется SQL-запрос к базе данных. После выполнения оператора SQL пользователь получает результат выборки. В настоящее время начато создание Web-узла на основе данного приложения.

Основная страница Web-узла содержит HTML форму панели выборки, аналогичную имеющейся в приложении. Связь Web-приложения с базой данных согласований осуществляется через ODBC-процессор. По результату выборки генерируется файл в формате HTML, который отправляется на выходную страницу Web-браузера.

Для интерактивного взаимодействия с базой данных используются апплеты Java. Для реализации взаимодействия «клиент-сервер» используются сервлеты.

3. Заключение

Предложенная модель удовлетворяет, как представляется, требованиям а)-в), сформулированным в начале статьи (описание согласования у Иомдина, несмотря на формализованный характер, этим требованиям не удовлетворяет). Модель позволяет объяснить целый ряд конструкций. Выражается надежда, что описанные механизмы соответствуют реальным когнитивным процессам. Создана база данных согласования на русском языке. Она данных может быть использована как для исследовательских целей, так и в учебном процессе. База данных создана на основе описания согласования в монографии [2] и нашей модели, и сама хорошо согласована с базай данных Суррея.

Литература

1. Surrey Database of Agreement. www.smg.suurey.ac.uk.
2. Иомдин Л.Л. Автоматическая обработка текста на естественном языке: модель согласования. М.: Наука, 1990.