

Смысловая неполнота в формуле слабой связи¹

Н.Н.Леонтьева

Введение

Известно, что основным источником нежелательного размножения связей при автоматическом синтаксическом анализе предложений служат так называемые "слабые связи", то есть связи, не предсказанные и не востребованные какими-либо словами в составе предложения. Именно такие случаи вызывают необходимость подключения более содержательного, чем синтаксический, аппарата, - им естественно является аппарат семантического уровня анализа. При семантической интерпретации невостребованные единицы (С) можно отобразить неполной триадой вида $P(? , C)$ либо $P(C, ?)$, если семантикой самого С предсказывается смысловое отношение Р, но неизвестен другой член этого отношения. Такие формулы - эксплицитное выражение локальной неполноты в семантической структуре предложения. Степень неполноты еще глубже, если невостребованная единица С не предсказывает даже имени отношения, которым она может быть привязана к какому-либо узлу структуры. Обозначим такую неполноту формулой $P(C, ?)$. Подобная запись будет сопоставлена, например, неузнанному "слову" типа "РФФИ:99-06-12657", особенно если оно подано в тексте отдельным абзацем. Неполные формулы в первичной семантической структуре являются сигналом того, что необходимо восстановление локально отсутствующих или неопределенных элементов формулы слабой связи.

Иногда структурный контекст таков, что и без обращения к семантике синтаксическая связь устанавливается однозначно (например, если кандидат на главный член - единственный), но, во-первых, остается неопределенность смысловой интерпретации связи, а во-вторых, необходимо решение этой задачи в общем виде, для всех остальных случаев, а их большинство (слабых групп несколько, кандидатов на главный член тоже несколько, к тому же кандидаты на соединение отношением Р сами имеют по несколько альтернативных значений). Поэтому обращение к аппарату семантики неизбежно.

Заполнение неполных формул может производиться или в пределах и из лексического материала того же предложения, или оно требует выхода за пределы предложения, в структуру целого текста. Механизм восстановления в том и другом случае один и тот же, - он апеллирует к Грамматике смысловых отношений. Грамматика предписывает нормативную синтагматику смысловых элементов (сведения о том, какие семантические характеристики - СХ - допускаются на местах семантических отношений - СемО) и их парадигматику,

определяющую степень сходства и возможность замещений-совмещений на множестве СемО и на множестве СХ.

Сложность правил интерпретации существенным образом зависит от того, какой результат поставляет синтаксический анализ, а именно: полный он или частичный, синтаксис составляющих, или деревьев подчинения, или их комбинации, синтаксические связи устанавливаются между словами или сначала строятся синтаксические группы, наконец, связи между ними именованные или просто фиксируется факт связи. В данной статье попутно с основной темой обсуждается вопрос о том, какая исходная структура необходима и достаточна для работы семантического компонента в нашей модели автоматического понимания текста. Обосновывается "взрывной" характер перехода от синтаксических групп к семантическим узлам.

1. Обоснование задачи

Мы возвращаемся к теме смысловой неполноты, поскольку она самым тесным образом связана с семантическим анализом естественных текстов. Предложенные впервые в [1] и следующих работах автора состав Смысловой грамматики и алгоритмы ее использования при анализе естественных текстов были опробованы и подтверждены несколькими реализациями семантического анализа в прикладных системах. В данной статье продолжаются наблюдения над смысловой неполнотой высказываний и текстов, - они будут сформулированы ниже с учетом уточнений, которые нам удалось внести в определения основных единиц и крупных составляющих семантического компонента.

Почему важно фиксировать смысловую неполноту? Прежде чем ответить на этот вопрос, напомним ту задачу, в рамках которой ведется исследование видов смысловой неполноты на данном этапе. Конечной целью анализа текста объявляется автоматическое построение Базы текстовых фактов (БТФ). Метод - создание (на стадии анализа) множества лингвистических структур, отображающих разные аспекты состава и содержания текста, затем синтез единиц, наиболее "важных" для данного текста, или наиболее полно отражающих содержание этого текста.

Из множества лингвистических структур, сопоставляемых целому тексту, отметим Семантическое пространство (СемПрост) текста, Ситуативное представление текста (СитП) и окончательное Информационное представление (ИнфП) текста. Их объединяет общее имя - Семантическое представление (СемП).

Важными среди промежуточных структур (или проекций) текста являются Терминологические представления, которые заведомо неполно передают содержание текста (например, в них отсутствуют связи), но во-первых, они легче других структур реализуемы, а во-вторых, они сразу могут начинать "работать" на больших массивах (см. статьи Н.В.Лукашевич и Б.В.Доброва в нескольких сборниках материалов ДИАЛОГа). В настоящее время поступила в сеть Интернет версия Университетской информационно-поисковой системы "РОССИЯ" (www.cir.ru), созданная коллективом Центра информационных исследований (ЦИИ), базирующимся в НИВЦ МГУ. Экспериментальные работы по семантическому анализу политических текстов также ведутся в рамках ЦИИ.

Из "малых" единиц, составляющих любые лингвистические структуры текста, назовем семантические узлы (сокращенно СемУ, или СемУзлы) и упомянутые СемО между ними. Для сформулированной выше задачи построения БТФ наиболее важными будут СемУзлы, обозначающие ОБЪЕКТЫ, и СемУзлы, обозначающие ДЕЙСТВИЯ либо СИТУАЦИИ (СИТ), связывающие эти объекты; а из множества выявленных семантических отношений, естественно, наиболее значимы единицы, обозначающие РОЛИ, выполняемые каждым объектом в данной ситуации.

Такова упрощенная схема любого сообщения, она лежит в основе многих грамматик (традиционных, а также ситуативных, ролевых и др. разновидностях грамматики в системах искусственного интеллекта), каждая система вольна называть эти единицы по-своему (например, имя действия называют отношением и т.п.), - важно лишь, чтобы они были непротиворечиво объяснены в составе одной модели или системы. В нашей модели при отображении в семантическом представлении единицы категорий СИТ и ОБЪЕКТ являются узлами (СемУ), а РОЛИ - семантические отношения (СемО) между ними:

СемУ_СИТ ----- > СемУ_ОБЪЕКТ

СемО (=РОЛЬ)

Сами семантические узлы как правило не даются в тексте в готовом виде, - их надо собрать в ходе анализа всего текста. В меньшей мере, но это же относится и к имени смысловой роли: часто его можно сформулировать только после применения ряда выводов на множестве смысловых отношений. Это объясняется как минимум двумя причинами: 1. главное слово семантического узла НЕ обязательно совпадает с главным словом синтаксической группы; 2. семантические единицы и целые структуры образуются не перекодировкой, а способом "взрыва", так как построение структур только с полноценными единицами невозможно без потерь, или "сброса", части информации.

Если при семантическом анализе удалось собрать полноценный узел СИТ, найти участников ситуации (ОБЪЕКТы) и определить их РОЛи, будем считать, что сформирована единица типа "текстовый факт" (ТФ), и он объявляется полномочным представителем данного текста во внешней среде. В частности, его можно внести в состав чьей-то индивидуальной Базы знаний (естественно, со ссылкой на источник получения этого фрагмента знания).

Вернемся к неполноте. Фиксируя порцию "Знания", нужно эксплицитно выразить и то, ЧТО данный текст "не знает" (об упомянутых ситуации или объекте). Это проявление смысловой неполноты в масштабе целого текста, что фиксируется в окончательной структуре ТФ в явном виде: ведь незнание составляет существенную часть знания. (Конечно, причиной "белых пятен" в структуре Знания может быть и недостаточно сильный аппарат анализа, но мы такие объяснения оставляем пока в стороне). Поверхностное выражение отсутствующего члена какого-либо отношения вызывает вопрос "Кто пришел?", "Куда пошел?", "Что спровоцировал?", "Зачем защищают?" и т.п.

К структуре связного текста мы приходим после анализа каждого отдельного предложения. Без фиксации неполноты и неопределенности, возникающих на локальных участках, невозможен выход за пределы предложения к целому тексту, так как смысловые опущения ведут себя подобно местоимениям: они отсылают к другим частям того же текста, из которых может быть восстановлен опущенный фрагмент, обеспечивая тем самым главное свойство текста, называемое связностью. Говоря "локальные", мы имеем в виду прежде всего неполноту семантических узлов типа ОБ или СИТ в составе отдельных высказываний (обычно предложений), которые сами являются узлами типа СИТ. Семантическое представление (СемП) отдельного предложения всегда содержит неполноту, т.е. вопросы, ответ на которые надо искать в СемПах следующих за ним или предшествующих отрезков текста.

2. Неполнота при построении семантических узлов

Основной вид смысловой неполноты - локальная неполнота СемУзла. Семантические узлы, в отличие от узлов синтаксической структуры (эти последние - обычно слова или

словоформы), являются довольно сложными единицами: в семантическом графе текста это подграфы, т.е. молекулы, образованные из атомов-слов, соединенных определенными смысловыми (или формальными, квазисмысловыми) отношениями. Отношения эти берутся в основном из имен валентностей слов, входящих в семантический узел.

В соответствии с определенной выше задачей анализа мы стремимся построить максимально полноценные СемУзлы, это значит такие, которые имеют хорошие оценки по нескольким параметрам, главные из них:

- а. полнота (заполненность всех внутренних валентностей),
- б. корректность (они построены в соответствии с правилами заданной смысловой грамматики),
- в. участие в ролевой структуре, входящей в основную тему текста,
- г. наличие внешнего денотата (или сравнимость с заданными вне данного текста объектами).

Возможны разные степени нарушения этих критериев в составе строящихся узлов. Каждый из несоответствий создает свой вид неполноты. Рассмотрим несколько простых примеров.

(1) Защита Белого Дома.

Это высказывание неполно по нескольким причинам. Оно неоднозначно: Белый Дом можно интерпретировать как АГЕНТа защиты (если в ходе семантического анализа этот термин получил семантическую характеристику $SX=ОРГ, ОДУШ$) или как ОБЪЕКТ (то, что защищают). Такая неоднозначность - свидетельство валентной неполноты; в данном примере при выборе первой интерпретации будет отсутствовать ОБЪЕКТ, а при второй отсутствующим окажется АГЕНТ. В том и другом случаях незаполненной остается валентность КОНТРАГЕНТ (от кого защищают). Все три валентности относятся к обязательным ситуативным связям, и отсутствие любой из них существенно снижает оценку данного выражения и построенного для него СемУ по полноте.

Участие данного СемУ в ролевой структуре главной темы документа (параметр в.) не определено, пока не установлены связи с включающим текстом. Если это заголовок, то оценка по параметру в. будет самой высокой, если мы рассматриваем его как изолированное высказывание (т.е. непонятно, в связи с чем говорится о защите Белого Дома), оценка будет минимальной.

Как влияет на оценку полноты или неполноты СемУ "внешний денотат" (пункт г.)? "Внешним денотатом" в нашей (вербальной) модели называется перечень (или иной вид перечисления/ задания, тоже в вербальном виде) объектов, относящихся к той же предметной области, что и текст. Частный случай такого задания являют собой Тезаурусы или Базы данных. Если построенный СемУ совпадает с одним из терминов Тезауруса (например, СемУ_Белый_Дом нашелся в политическом Тезаурусе), этот узел получает большой денотативный вес, или высокую оценку полноты по параметру г. По сути дела, это полнота информации о данной единице с точки зрения внешней системы. "Узнать" построенную анализом единицу означает получить доступ к дополнительной информации о ней, что позволит расширить СемУзел. Так, СемУ с $SX=ФИО$ (напр., Г.Н.Селезнев) может быть расширен добавлением должности ($SX=ДОЛЖ$, например, Председатель Госдумы).

Возможно узнавание семантической единицы по заданному извне образцу, напр., "12.06.1989" - это полный СемУзел типа ДАТА, а "121.06.1989", не совпадающий со схемой даты, - неопределенный СемУ, которому может быть приписано несколько возможных

интерпретаций, тем самым он семантически неполон (в нем недостает информации, необходимой для его отождествления).

Иногда дополнительная информация из внешних источников не уточняет, а "размывает" информацию об узле. Так, единица "Белый Дом" может получить как минимум два альтернативных уточнения - (ЛОК=Вашингтон или ЛОК=Москва). Это означает, что в СемП добавляются вопросы (т.е. неполнота).

3. Первичная семантическая интерпретация "сильных" связей как источник появления некорректных (неполных) семантических формул

Трудно было бы ожидать, что после применения самого простого правила семантической интерпретации синтаксических связей ($a-r-b \Rightarrow A-R-B$) мы сразу получаем корректные семантические формулы, соответствующие заданной смысловой грамматике. Во-первых, напомним, что a и b - это слова, а A и B - семантические узлы, представляющие собой, как правило, словосочетания, или группы слов, или семантические подграфы. Уже поэтому лобовое переименование связи синтаксической в семантическую приводит часто к искажению смысла языкового выражения. Рассмотрим примеры.

(2) "Одну столовую ложку травы нужно залить двумя стаканами кипящей воды".

В этом тестовом и неоднократно приводившемся нами примере прямая интерпретация даст выражение ЛОЖКУ ЗАЛИТЬ СТАКАНАМИ, с двумя некорректными формулами: КОНечной_Точкой (куда?) ЗАЛИВания должно быть ВЕЩЕСТВО, а не ПРЕДМЕТ (ложка), и ОБЪЕКТом (что или чем заливают?) действия является ВЕЩЕСТВО + ЖИДКОЕ, но не ПРЕДМЕТ (стаканы). Если запретить построение формул с нарушением правил смысловой грамматики, образуются две неполные формулы. Требуется либо скорректировать такое первичное СемП, либо поменять порядок действий, т.е. сначала построить правильные семантические узлы, а затем интерпретировать связь. Правильными в данном примере будут узлы, в которых синтаксически зависимые слова станут семантически сильными: ТРАВА В КОЛИЧЕСТВЕ одна столовая ложка ЗАЛИВАЕТСЯ ВОДОЙ кипяченой В КОЛИЧЕСТВЕ два стакана. Пример (2) иллюстрирует семантически неустойчивую синтаксическую группу (вернее, две такие именные группы). В нем же и предикатная группа является семантически неустойчивой, так как ее главное слово ("нужно") будет подчиненным в составе семантического узла: МОДАЛЬНОСТЬ(нужно,залить). Кроме количественных и модальных, к тому же разряду относятся СТАДИЙНЫЕ, параметрические и некоторые другие конструкции. Все они требуют применения правил смысловой коррекции, меняющей иерархию единиц в составе узла.

Другой пример:

(3) "Едва ли не главной целью путчистов был телецентр."

Прямая интерпретация синтаксических связей привела бы к формуле *ЦЕЛЬ(телецентр,путчисты). СемП этого высказывания неполно дважды, так как на первом и втором местах отношения ЦЕЛЬ(,) должны быть СемУ "ДЕЙСТВИЯ" или "СИТУАЦИИ". Это противоречие грамматике смысловых отношений порождает дважды неполную формулу ЦЕЛЬ(?А,?В), где ОБЪЕКТ(телецентр,А) и АГЕНТ(путчисты,В). Цель может быть восстановлена как захват или иные ДЕЙСТВИЯ (?А) с телецентром, но не сам телецентр, а вторым, семантически главным членом отношения цели является не лицо (путчисты), а действие или ситуация, активным деятелем (АГЕНТОМ) которых они являются. Аналогично ведет себя смысловое отношение ПРИЧИНА(,), хотя на поверхностном уровне оно тоже

допускает ЛИЦО вместо ДЕЙСТВИЯ, ср. "Маша опоздала из-за Ивана":

*ПРИЧИНА(Иван,опоздание Маши) => ПРИЧИНА(?А,опоздание), где АКТАНТ(Иван,А).

4. Неполнота в формуле слабой связи

Рассмотрим примеры с интерпретацией слабой синтаксической связи.

(4) "Реализация природного газа конечным потребителям без участия газосбытовых организаций (перепродавцов) непосредственно поставщиками (предприятиями и объединениями по транспорту и поставкам газа) в порядке, предусмотренном в п.2 настоящего постановления, осуществляется..

(5) "Реализация на внутреннем рынке Российской Федерации природного газа предприятиями и организациями, независимо от ведомственной принадлежности и форм собственности (включая биржи), по ценам, превышающим установленный предельный размер государственных регулируемых цен, является ...". (Из Собрания актов Президента и Правительства РФ).

В этих двух типичных для жанра документах больше слабых связей, чем сильных. У слова "реализация" синтаксически сильными можно признать лишь Род.п. и Тв.п. (Р. чего и кем, обозначающие семантические ОБЪЕКТ и СУБЪЕКТ реализации). Привязка остальных групп может проходить на семантическом уровне, в соответствии с семантикой самого слова или значения связующего элемента. На первом шаге интерпретации слабоуправляемые группы перейдут в неполную формулу: группа "конечным потребителям" - в АДРЕСАТ(конечные потребители,?В) - здесь именование связи идет от значения дательного падежа. Группа "без участия..." образует формулы АКТАНТ(газосбытовые организации,?) и МОДАЛЬНОСТЬ(ОТСУТСТВИЕ,АКТАНТ). Группа "на внутреннем рынке Российской Федерации" порождает неполную формулу СФЕРА(рынок,?). Группа "по ценам, превышающим..." перейдет в формулу В_СООТВЕТСТВИИ (?СемУ_цены).

Рассмотрение именно слабых, разнообразных в поверхностном выражении связей позволяет тоньше и точнее описать семантику слова, для которого эти синтаксически слабые связи являются сильными семантическими валентностями. Так, к семантически сильным валентностям слова "реализация" можно отнести еще СФЕРУ(Д,Р.): 'Где, в какой сфере реализуется?' и УСЛОВИЕ(Е,Р.), но без задания конкретных способов выражения этой связи. В этом случае в предложении (4) валентность СФЕРА(,) будет "погашена" слабой группой АДРЕСАТ(,), поскольку второе является частным случаем первого. Заметим, что в случае, если в поле синонимов для слова "реализация" указано слово "продажа", у которого есть сильная валентность на адресата, присоединение группы "конечные потребители" к первому будет считаться удовлетворением этой валентности, а не ее "гашением". Формула АКТАНТ(газосбытовые организации,?) /отвлекаясь пока от ее отрицательной модальности/ может быть формально добавлена и к СУБЪЕКТУ, и к АДРЕСАТУ, но по позиционному критерию безусловное предпочтение отдается первому.

В предложении (5) мы находим две слабоуправляемые группы, из них СФЕРА(,) удовлетворяет одноименную валентность слова "реализация", а группа В_СООТВ(Р,цены) гасит валентность УСЛОВИЕ(Е,Р.).

Слабоуправляемые группы - это языковой материал, прежде всего требующий перехода на семантический уровень анализа, функции которого не ограничиваются рамками предложения. Слабые группы часто образуют семантически сильный узел. Напомним рассказ И.Бунина "Господин из Сан-Франциско". Слово "господин" имеет сильную смысловую

валентность ИДЕНТификация(? ,господин), удовлетворяется она отношением ИМЯ(? А,господин). Но на протяжении всего рассказа имя ни разу не появляется, вместо него всегда ИСХодная_Точка(Сан-Франциско,господин). В окончательном СемП вся эта формула, "гасящая" обязательную валентность, будет объявлена идентификатором "господина": ИДЕНТ(ИСХ_Т(С.-Ф.,г-н),г-н).

5. Неполнота изолированного высказывания

Семантическое представление вопросов, заданных на естественном языке, содержит законную, декларируемую неполноту. "Что случилось?" => СОДЕРЖ(? ,случиться); "Кто виноват?" => ПАЦИЕНС(? ,вина); "Куда пошел?" => КОНЕЧ_ТОЧ(? ,пойти); "Что пересек?" => ПРОМЕЖУТ_ТОЧ(? ,пересечь) и т.д. Неполная формула вопроса, обращенного к тексту, служит тем оператором, который должен сформировать как можно более полный ответ из материала СемП текста.

Как правило содержат неполноту заголовки к текстам. Существует большой разброс в типах заголовков. Иногда в заголовок сразу выносятся действующие лица, ситуации и объекты ("...как один мужик двух генералов прокормил"; "О переводе с угля на газ Тульского хладокомбината"; "Сотрудничество развивается"), анализ текста только добавляет к ним новые свойства или заполняет объявленные валентности. Хуже обстоит дело с анализом заголовков, содержащих многозначные слова или "обманную" модель управления ("Перевод денег"; "Базы данных о переводах на магнитных лентах"; "План Лаборатории машинного перевода с французского языка на 1999 год").

Локальная неполнота в заголовке часто объясняется слишком абстрактным характером заголовка ("Тише едешь - дальше будешь"), когда не удается провести от него никаких смысловых отношений к материалу текста. Неоднозначности, игра слов, метафорические переносы и другие художественные приемы делают автоматический анализ такого материала мало реальным.

Более всего пестрит неправильностями, где превышена степень допустимой неполноты, устная речь, особенно диалогическая речь, выступления на собраниях, в парламенте, в вопросах и ответах...

"Дать от микрофонов по порядку ведения".

"Вы по мотивам или по повестке?"

Интерпретация подобных высказываний должна опираться на списки речевых или специальных шаблонов, где значения предлогов и других лексических и грамматических средств связи не должны интерпретироваться.

(Неправильная) подмена сильного синтаксического управления моделью от другого слова (чаще всего в выступлениях депутатов) образует оторвавшиеся слабые группы:

*Он своими глазами увидел О ТОМ, что там творится (Зюганов)

*Отсюда видно О ТОМ, что происходит в системе образования

*Я хотел бы убедить депутатов О ТОМ, что в России началось новое время

*Предлагаю решить ПО ПОВОДУ О ТОМ, что ...

*То, что касается ОБ ЭТОМ ДЕЛЕ...

*Нужно попугать их (демократов) О ТОМ, ... (вместо ЧЕМ)

*Бандиты убеждают друг друга О ТОМ, что нечего бояться.

Везде оторвавшаяся (т.е. не присоединившаяся по правилам грамматики) слабая группа является семантически сильной, так как способ ее присоединения к предложению однозначно интерпретируется как СОДЕРЖАНИЕ(А,?В), где на место А встанет все последующее утверждение. Нахождение того СемУ (?В), к которому присоединится эта формально не затребованная группа, тоже не составит труда: это тот главный предикат, у которого осталась незаполненной валентность СОДЕРЖАНИЕ(?А,В). Так семантический анализ, имеющий дело с неполнотой, справляется с дефектам построения предложений, что вероятно, не учит грамотности депутатов, но зато развивает способности систем искусственного интеллекта.

Пример (2), кроме неполноты двух СемУзлов, иллюстрирует (неочевидную) смысловую неполноту всего высказывания. Ее образует модальный СемУзел "нужно залить", так как СемУ, у которого СХ=ДЕЙСТВИЕ в МОДАЛЬНОСТИ "НЕОБХОДИМО", приобретает валентность ЦЕЛЬ(?СемУ): "Зачем нужно залить что-то куда-то? Что дальше делать с полученным результатом?" Если идущий за этим текст отвечает на вопросы, мы запишем в Базу знаний ответы, если же высказывание остается изолированным, в базу знаний нужно поместить неполную формулу либо вообще изъять данное утверждение как неполноценное.

Сферой действия семантического анализа в нашем подходе является целый законченный текст, из которого можно извлечь "полезные" утверждения, заслуживающие их помещения в некую создаваемую базу знаний с указанием автора утверждения, массива и других обязательных атрибутов нормального текста. В рамках такой установки сведения о неполноте отдельного знания, высказывания, целого текста, массива становятся осмысленными. Напротив, нам трудно обсуждать проблемы семантического анализа на примере чисто лингвистических иллюстраций типа "Джон любит Мери" или "Джон - холостяк". Единственная валентность, которую можно приписать таким высказываниям - "Ну и что из этого?" или "Что Вы хотите этим сказать?"

6. От синтаксических групп к семантическим узлам

(уроки реализации)

В нашем подходе синтаксическое представление (СинП) предложений рассматривается лишь как ступенька к построению полноценных семантических единиц - узлов и связей. Поскольку СемУзлы в общем случае сложные единицы (словосочетания), в качестве исходной синтаксической модели был выбран синтаксис групп А.В.Гладкого [2] ; программная реализация была выполнена Н.Ф.Суцанской [3]. Понятие синтаксической группы пришлось расширить (в частности, включить определения в состав именных и предложно-именных групп и др.). Мы не предъявляли жестких требований к результату синтаксического анализа (СинАн) как отправной точке для семантической интерпретации: допускалось, что СинП может и не быть правильной, однозначной, полной и т.д. структурой, поскольку за ним следует более мощный семантический анализ. Однако логика реализации постепенно привела к масштабу "полного" синтаксического анализа. Такая сложная система требовала

сильной команды программистов и лингвистов, каковой мы не имели. Отъезд ведущего (и единственного ответственного за синтаксис) программиста привел в нашем случае к тому, что преемники не смогли развивать программы.

Почему так быстро нарастает сложность СинАн? Во-первых, начиная с какого-то момента (построение связей в соответствии с моделью управления) синтаксис все больше требует поддержки со стороны семантики, что приводит к размыванию строгих грамматических критериев, то есть к выходу за рамки чистого синтаксиса, к сложному синтактико-семантическому анализу. Таков ход практически всех известных нам развитых прикладных систем: это системы семейства TAUM (TAUM-METEO, TAUM-AVIATION и др.), не получившая развития система EUROTRA и др. [4], а также системы, основанные на модели СМЫСЛ-ТЕКСТ [5]. Их роднит осторожное использование семантики – как аппарата, дополняющего синтаксис. В результате основой СемП остается синтаксическая структура, структура связей между СЛОВАМИ. Семантическая интерпретация, которая состоит в основном в уточнении значений слов (стол1, стол2 и т.д.) и переименовании связей (в глубинно-синтаксические или так называемые "падежные"), постепенно преобразует синтаксическое дерево в семантизированный граф. Этот процесс остается "эволюционным" до тех пор, пока он протекает внутри каждого отдельного предложения текста.

Переход к структуре целого текста неизбежно требует ломки структуры, "взрыва": проведение межфразовых связей не только нарушает целостность синтаксической структуры, но и заставляет пересмотреть состав и границы некоторых построенных синтаксисом единиц (как было показано выше). Сигнал о необходимости коррекции (пересмотра иерархий внутри группы) поступает от первичной семантической структуры, более конкретно, от появившихся в ней "дыр" (или локальной смысловой неполноты). Восстановление референтов неполных формул тоже может привести к построению единиц, отличных как от данных синтаксисом, так и от тех, которые были образованы в соответствии с первичными валентностями, заданными в семантическом словаре.

Поэтому в новом варианте СинАн было решено провести более строгое разделение синтаксического и семантического компонентов. За синтаксисом оставлена задача сборки ГРУПП составляющих (для группы формируется полная грамматическая характеристика и приписывается оценка достоверности связей), а начиная с вызова семантического словаря анализ передается семантическому компоненту.

Исследование явлений, происходящих на границе синтаксического и семантического компонентов, продолжается в рамках НИВЦ МГУ (Университетская информационная система "РОССИЯ"). Реализация одного из вариантов стыковки результатов СинАн с аппаратом СемАн, равно как и реализация упрощенного СинАн, осуществляется сейчас в рамках другой системы (см. статью А.Сокирко в наст. сборнике).

Заключение

Семантическая интерпретация строит единицы другой природы, чем синтаксический анализ, она часто сопровождается образованием пустот, "дыр" в структуре, которые эквивалентны вопросу, возникающему на данном участке текста. Наиболее серьезно такие вопросы ставит проблема присоединения слабоуправляемых синтаксических групп: необходимо вскрыть семантику самой группы (т.е. построить СХ всего СемУзла из СХ составляющих его слов), определить значение средства связи (семантическое отношение, передаваемое предлогом, союзом и др.), построить СХ возможных кандидатов на второй член отношения. Гипотеза о

семантике такой формулы слабой связи должна быть проверена на непротиворечие другим формулам текста, в которые входят СемУзлы, участвующие в интерпретации слабой связи.

Естественный текст - это всегда диалог: каждое предложение в нем и утверждает нечто, и ставит вопросы, и отвечает на вопросы, поставленные в других предложениях текста. Неполнота в тексте движет семантический анализ. Иногда за недостающими в текстах сведениями приходится обращаться к семантике массива (например, в массиве газетных статей, когда мы ищем референта для СемУзлов типа "Вчера", или "Ровно месяц назад", или "Сегодня рано утром"). И в этом случае, как при восстановлении СемУзлов в составе предложения или в масштабе целого текста, работает та же Грамматика смысловых отношений.

Смысловая неполнота, которая эксплицирована в виде вопросов (пустот, дыр в структурах), выполняет конструктивную роль: она используется для получения связной структуры целого текста. Анализ смысловой неполноты - это краеугольный камень семантического анализа связного текста.

Литература

Н.Н.Леонтьева. О смысловой неполноте текста (в связи с семантическим анализом) // Сб. Машинный перевод и прикладная лингвистика, вып. 11. - М., 1970.

Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. - М.: Наука, 1985.- 143 с.

Суцанская Н.Ф. Программный препроцессор для естественно-языковых интерфейсов: Автореф. дисс. к.т.н. - Киев: РИО ИК, 1989.

Искусственный интеллект. Справочник. В 3-х кн. М.: Радио и связь, 1990.

Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. Лингвистическое обеспечение системы ЭТАП-2: Наука, 1989. 295с.

1 Исследования смысловой неполноты в текстах официальных документов ведутся при поддержке Российского Фонда фундаментальных исследований (Грант РФФИ:99-06-80296).