SLAVICA HELSINGIENSIA 35
С любовью к слову
Festschrift in Honour of Professor Arto Mustajoki on the Occasion of his 60th Birthday
Ed. by Jouko Lindstedt et al.
Helsinki 2008
ISBN 978-952-10-5136-4 (paperback), ISBN 978-952-10-5137-1 (PDF), ISSN 0780-3281

Нина Николаевна Леонтьева
(Научно-исследовательский вычислительный центр Московского государственного университета им. М.В. Ломоносова)

Смысловая неполнота или неграмотность? (Где, когда и с кем *Лодку унесло ветром?* Взгляд прикладного лингвиста)

К юбилею Арто Мустайоки

A la recherche du sens perdu (по М.Прусту)

1. Почему надо исследовать смысловую неполноту

Имея дело с текстовыми корпусами, которые обрушиваются на нас потоками, мы не можем не пересмотреть роль многих лингвистических явлений в этом новом свете. Принципиально новым оказывается не только возможность, но и необходимость выхода за пределы предложения, а значит, за пределы моделей анализа, основанных на синтаксисе. Можно обратиться к тем семантическим явлениям, которые возникают на границах и даже далеко за границами отдельных синтаксических и синтактико-семантических структур. К таким явлениям относится и смысловая неполнота текста (СНТ) и предложения. В прикладных системах рассматривается только синтаксическая неполнота, или эллипсис, как наиболее очевидная причина и смысловой неполноты предложения, как дефект, а решение задачи восстановления недостающих частей структуры откладывается обычно «на потом». Но СНТ проявляется не только в виде эллиптичных предложений; развитые прикладные системы, основанные на полном лингвистическом анализе (именно их мы обозначаем далее словом **Системы**), сталкиваются с неполнотой данных на всех уровнях. Дадим самое общее определение СНТ: в анализируемом фрагменте отсутствуют (опущены) такие части, без которых нельзя построить «полную и правильную» структуру данного уровня. Если на синтаксическом уровне свойства «полной и правильной» структуры достаточно очевидны, то в семантике наблюдается большое разнообразие субъективных определений того, что такое правильная семантическая структура (или СемП).

Простор возможных субъективных толкований явления смысловой неполноты (как и ряда других семантических проблем) может быть несколько ограничен именно **прикладной** точкой зрения, потому что прикладной лингвист должен предъявлять, хотя бы в модельном виде, те сущности и компоненты Системы, *относительно* которых (сравнением с которыми) можно обнаружить неполноту. В данной статье подчеркнут аспект **относительности** смысловой неполноты.

Реальные естественные тексты (ЕТ) и текстовые корпуса (ТК) являют собой опытное поле, в пределах которого можно будет исследовать неотъемлемые свойства действительно **целого текста**. К ним относятся а) избыточность, или дублирование элементов смысла текста, провоцируемая стремлением к надежности сообщения, б) компрессия как прием устранения избыточности, в) локальная неполнота как следствие эллипсиса или иной смысловой компрессии, г) связность, на которую опирается не только выявление всех названных свойств, но и восстановление недостающих частей полного («идеального») состава текста. При описании любого из названных явлений необходимо учитывать весь этот комплекс. **Позитивная роль** СНТ состоит в том, что при ее явной фиксации в некоторой семантической структуре отображающие неполноту символы обнажают участки **связности** текста. СНТ можно считать «краеугольным камнем» автоматического понимания текста (АПТ).

Корпуса нужны в нашем случае и для того, чтобы сортировать выявленную при анализе текста локальную неполноту, и для того, чтобы проверять по ним правильность и эффективность тех решений, которые мы принимаем для экспликации и дальнейшего сжатия содержания текста. Это даст возможность вычислять затем **степень неполноты** или **неопределенности** целого текста, равно как и **меру связности** – ведь все это составляющие качественной оценки информационных свойств текста. Побочным результатом анализа на неполноту может быть программа стилистического редактирования создаваемого текста, которая может задать вопрос типа «Не хочет ли автор уточнить Причину или Время описываемого события?» и т.п.

2. Неполнота в корпусе

Смысловой неполноты в корпусах текстов и в любом наугад взятом тексте намного больше, чем кажется на первый взгляд. Но при поиске какого-то сложного лингвистического явления в ТК мы можем задать лишь лексему, или грамматическую характеристику, или синтаксическую конструкцию, или интересующие нас семантические характеристики лексем, слов и словосочетаний, или даже разные комбинации упомянутых признаков. А что можно задать, если мы ищем нечто про-

пущенное, невыраженное, пустоту, нуль? Как собрать массив фраз, содержащих смысловую компрессию? – ведь такой материал внешне ничем не выделяется: в корпусе текстов неполнота не видна и даже **неуловима**. Мы не можем запросить «Выдать все семантически неполные предложения из такого-то текста или корпуса». Если решить, что семантически неполными являются все односоставные предложения (не содержащие подлежащее либо сказуемое), то видимо, даже и по синтаксически размеченному корпусу проблематично их найти, так как построенные синтаксические структуры не снабжаются такими характеристиками, как тип предложения. Видимо, нетрудно задать алгоритм выбора из потока текстов тех предложений, которые состоят из одной лексемы или из одной предложной группы, то есть заведомо неполных и семантически, снабдив эти примеры даже контекстом заданного размера для дальнейшего изучения. Но это лишь малая часть того, что неполно по смыслу: ведь и синтаксически полная и формально правильная фраза часто семантически неполна (*В России можно получить 7 лет за то, за что в Европе получают Нобелевскую премию*, *сказал посол*) или содержит минимум смысла (*В этом направлении уже сделаны определенные шаги*). Наконец, можно собрать массив газетных заголовков. Но нас все они интересуют лишь как первичный материал, содержащий с большой вероятностью и смысловую неполноту.

Да, возможности обращения к ТК в поисках неполноты крайне ограничены, однако не совсем верно утверждение, что в корпусе неполнота *не видна*. На этапах первичного (графематического) анализа Система находит много внешних, поверхностных признаков неполноты как результатов сжатия. Это и разные виды сокращений: Ин-т, Ун-т, Пр-т, ул., Моск., в %-м отношении, 23-й, 459-ом, и другие – стандартные и нестандартные. Это и разные аббревиатуры, если им не дана немедленная (в той же фразе) расшифровка, как в *ПАСЕ – Парламентская Ассамблея Совета Европы*. Это и словосочетания: *и т.д., и т.п., и др.; и пр., и пр., и пр.* и подобные им; они свидетельствуют о логической незавершенности высказывания; и они же требуют разрешения омонимии точки (конец фразы или знак сокращения).

О неполноте (незаконченности) отдельного предложения могут свидетельствовать также знаки многоточия, двоеточия, тире, запятой, точки с запятой или вообще отсутствие какого-либо знака препинания на конце (ведь в Системе конец фразы отмечен знаком абзаца). Например, *Я далек от мысли…* (Черномырдин).

Из других графических признаков неполноты назовем отсутствие закрывающих скобки и кавычки при наличии открывающих их пар (или наоборот: есть только закрывающий член парного знака); редактор Word подчеркивает их как возможный сигнал ошибки. Пропуск номера пункта в цепочке перечислений (1, 2, 4, 5,…) – это уже гло-

бальный признак ошибки, в отличие от локальных, названных выше. Чтобы его обнаружить, нужна программа счета пунктов перечисления (п/п) по тексту.

Все названные случаи относятся еще к графематическому, нулевому уровню, это обнаружение чисто внешних признаков какой-то неполноты. Причина же неполноты – сжатие за счет сокращения избыточной информации, легко восстанавливаемой пользователем. Большинство таких первичных текстовых проблем решается заданием эмпирических правил и списков. Для Системы АПТ, проводящей полный лингвистический анализ, включая уровень семантики, это создает определенные проблемы. А непрерывные фрагменты текста, не нашедшие решения, Система может временно «инкапсулировать» как неопознанные лингвистические объекты (НЛО).

Наконец, из ТК можно выловить большой класс «полуместоименных и местоименных» слов, семантика которых свидетельствуют о лексической и тем самым о семантической неполноте. При семантическом анализе нужно найти их текстовый «референт»: *так*, *он*, *аналогично*, *делать*, а также многие другие обобщающие лексемы, которые замещают имена или описания конкретных объектов. Приведу цитату из книги Н.Д. Арутюновой (1988, 102): «Такие имена, как *факт*, *событие*, *ситуация* и др., постоянно вступают в анафорические и катафорические отношения в тексте, замещая или вводя пропозициональные конструкции, номинализации или даже целые фрагменты текста».

3. Неполнота и сжатие в Анализаторах и Генераторах

Из свойства неуловимости СН в корпусе можно сделать первый поверхностный вывод о том, что изучать смысловую неполноту по корпусам сейчас практически мало реально, по крайней мере, до тех пор, пока не предложен какой-то механизм семантической интерпретации связей. Остается умозрительный способ моделирования семантического анализа предложения и вслед за ним (обязательно) целого текста. Главным средством анализа при этом должен быть некоторый семантический метаязык, отличный от синтаксического, поскольку область действия неполноты, а главное, область, которая позволяет восстановить полный вид высказывания, часто лежит далеко за пределами отдельного предложения. И сами синтаксические и семантические структуры, и неполноту как свойство той и другой структур нужно исследовать отдельно, используя при этом разный аппарат на этих разных уровнях.

При семантическом анализе заданного текста цель состоит в том, чтобы доказать содержательную связность текста, а для этого потребуется восстановить фрагменты, без которых многие связи не будут фор-

мально выражены. Результат такого восстановления может быть отображён в некоторой вспомогательной структуре, где для каждой местоименной, полупустой, замещающей единицы указан ее текстовый референт в виде конкретной лексемы или фрагмента структуры. Например, РЕФ(*Иван Петрович,**он***). При этом неоднозначность восстановления первого члена связи РЕФ мы тоже относим к смысловой неполноте. Так, в примере Л.Н. Иорданской *Он подвел ее к кушетке и сел на нее* для последней лексемы (*она*-2) будут построены две омонимичные отсылочные связи:

РЕФ(*она*-1, *она*-2) и РЕФ(*кушетка, она*-2).

Референты заместителей *Он* и *она*-1 могут быть установлены на одном из первых шагов семантического анализа обращением к словарному описанию лексемы *сесть* и доказательством связи этого высказывания с предшествующими ему. Не найденный при анализе РЕФерент отображается в СемП знаком вопроса: «?». Ср. мнение А. Мустайоки, который утверждал, что неполнота высказывания проявляется в том, что в семантической структуре высказывания больше единиц, чем в поверхностном выражении. Это естественно, так как для формального восстановления связности нужно фиксировать межфразовые связи и выявлять имплицитные части, которых недостает для доказательства связности. Кроме того, и другие виды неполноты получают в промежуточном семантическом графе каждый свое отображение. На завершающей стадии анализа целого текста стоит задача минимизировать локальную (вре́менную) неполноту путем «замыкания» многих связей, обобщения и оценки самих неполных участков.

Восстановление «полного и правильного» вида записи, отражающей **содержание** целого текста в Системе АПТ, может руководствоваться двумя установками. Либо это установка на самое точное воспроизведение состава, акцентов, распределения самой неполноты и прочих деталей исходного текста по фразам (что нужно для абсолютно полного машинного перевода – МП), и тогда мы идем по пути классических лингвистических моделей класса «Смысл-Текст». Либо это установка на извлечение информационного содержания текста, ранжирование его на основное и второстепенное, устранение слишком детальных, или избыточных, или «непонятых» Системой частей и т.п., и тогда окончательное СемП строится по законам информационных объектов и структур. Его правильнее назвать информационным представлением (ИнфП), в котором «знание» включает и «незнание» (в нашем случае неполноту) как неотъемлемую часть (по Д.А. Поспелову). При вербализации ИнфП средствами ЕЯ неполнота получает вид естественного вопроса.

В генерируемом тексте надо вводить грамотную, правильную и даже обязательную неполноту, избегая того, что искажает языковые нормы либо приводит к неоднозначному или даже неправильному пониманию высказывания, хотя правилам сжатия и соответствует. Нужны две РАЗНЫЕ грамматики – для понимания и для воссоздания текста. Подчеркнем **несимметричность** этих двух процессов. Если при анализе мы имеем дело со стихийной неполнотой и интуитивно ощущаемыми автором текста нормами стилистики (стремлением к краткости и желанием избегать повторов), то при синтезе текста важно соблюдать правила корректного, или допустимого, не мешающего восприятию, сжатия; это касается больше сжатия научного текста. На этапе вербализации семантической структуры совсем не обязательно, чтобы сжатие проходило в тех же позициях, где были смысловые опущения в исходном тексте. Даже и в системах МП необходимо сжатие по законам выходного языка. Несимметричность описываемых процессов подчеркивается тем, что все более востребован перевод через базы знаний, то есть со сжатием исходного текста до такого вида, который соответствует поставленной задаче. Еще более важна минимизация текста и в других интеллектуальных системах – автоматическое реферирование, извлечение знаний количественного характера, ИПС, см. работы о смыслосохраняющем сжатии (Гиндин 1982, и другие). В любом варианте Система должна регистрировать каждый локальный случай «несогласия» с автором текста, выдавая сигнал об ошибке или неоднозначности хотя бы в промежуточной структуре; давать же реальную оценку неполноты и неопределенности лучше после завершения всего процесса работы с текстом.

4. Зависимость оценки неполноты от жанра

В корпусной лингвистике (КЛ) предлагается, а в каких-то корпусах уже принято снабжать тексты индексом жанра. Наиболее очевидные классификации по жанрам – это деление на поэзию и прозу, а сама проза может быть художественной, научно-популярной и научной или деловой. Уже такая простая классификация позволит различать стратегии анализа, даже автоматического: к поэзии нельзя применять такие жесткие правила синтаксиса, как к прозаическому повествованию, а словари не придется дополнять терминологическими тезаурусами, как того требует анализ научной литературы. Что касается семантики, то в аспекте неполноты самый большой материал поставляет жанр драматургии и записей разговорной речи, а также диалоги в составе повествования; в составе делового текста неполными и неопределенными оказываются заголовки (ЗГЛ), подзаголовки, подписи под рисунками и т.п. Система должна иметь регистр, настраивающийся на жанр, но при

отсутствии вообще текстового семантического анализа о таких тонкостях говорить пока рано. Ведь при определении полноты-неполноты высказывания или реплики нужно учитывать и цель автора текста: например, произвести художественный эффект нарочитой неоднозначностью (ср. *Каспаров теряет фигуру, но спасает партию* – ЗГЛ статьи; *Нет частной собственности* – лозунг; *Глаголы движения в воде* – ЗГЛ книги). Для пользователя важно, чтобы Система отметила сам факт неоднозначности. Не надо сразу относить его к неполноте, требующей исправления, как это желательно в деловых документах: ведь выбор и оценка художественных приемов остаются пока прерогативой человека.

Еще один жанр, богатый на явления локальной неполноты, – это тексты объявлений. Рассмотрим одно объявление в метро (приводится с точностью до знаков):

ЭКСПРЕСС КРЕДИТ
Москвичам и жителям Подмосковья
до 150000 руб.
БЕЗ СПРАВКИ О ДОХОДАХ, ЗАЛОГА И ПОРУЧИТЕЛЕЙ
Кредит выдается наличными в день обращения

Если этот текст будет анализировать Система, каждая строка будет сочтена отдельным простым предложением, а на семантическом уровне отдельным простым высказыванием. Каждое из них будет признано неполным, даже если игнорировать такой показатель неполноты, как отсутствие обязательного конечного знака препинания. (Замечу попутно, что в большинстве моделей синтаксиса составляющих и зависимостей знак препинания не включается в структуру, что создает трудности для последующей интерпретации: приходится обращаться к исходному виду фразы.) Эксплицируем наиболее очевидные смысловые опущения в виде вопросов к каждому из них. 1. *КРЕДИТ* – Кому и в каком размере? – по смысловым валентностям в словаре; 2. *Москвичам и жителям Подмосковья* – предполагаемый адресат какого действия? – по семантике дательного падежа; 3. *до 150000 руб* – это сумма является стоимостью чего? – как стандартная лексическая функция выражений, обозначающих деньги; и какой предел обозначен предлогом *до*? 4. *БЕЗ СПРАВКИ О ДОХОДАХ, ЗАЛОГА И ПОРУЧИТЕЛЕЙ* – это предложный оборот, оторванный от слова, им управляющего (О каком действии идет речь? Где утверждение?). К тому же он синтаксически неграмотный; 5. *Кредит выдается наличными в день обращения.* См. вопросы к лексеме *Кредит* + Кто *выдает* и кому? А также: Чье *обращение*, к кому? + Каково содержание *обращения*? – все это вопросы по смысловым валентностям лексем в словаре. (Подробнее о них см. в разделе 8).

Хотя все пять высказываний локально неполны, они образуют связный текст и на большинство вопросов сам же текст дает ответ – процедурой замыкания друг на друга (или взаимного насыщения) неполных формул. Текст в целом можно оценить (с учетом критериев, принятых для данного жанра) как удовлетворительный по признаку полноты, хотя одно предложение (4) содержит неграмотную неполноту (недопустима компрессия, состоящая в устранении повторяющегося предлога, особенно если во фразе есть еще и другие предлоги/падежи, претендующие на восстановление). Система должна сообщить о таком нарушении правила при анализе текста и предложить правильное сжатие конструкции при синтезе, а именно: «повторить предлог *БЕЗ: БЕЗ СПРАВКИ О ДОХОДАХ, БЕЗ ЗАЛОГА И (БЕЗ) ПОРУЧИТЕЛЕЙ*». К жанру объявлений при анализе применим и следующий уровень компрессии, которая приводит к построению фрагмента базы данных (БД), если текст содержит числовые параметры.

Остановлюсь на таком специфическом жанре, как лингвистическая литература. Это массивы текстов и разных словарей, которые изобилуют лингвистическими примерами, взятыми из текстов же или придуманными автором. К таким включенным в текст рассуждений или словарных определений «цитатам» надо относиться осторожно: при автоматической обработке уметь их вычленять, чтобы не применять к ним критерии оценок, справедливых для включающего текста. Например, из многочисленных употреблений и склонений фразы «Джон любит Мэри» нельзя делать вывод о пристрастиях автора на основе превышающих норму частотных характеристик этих трех лексем в статьях лингвистов. Нежелательно выносить эту фразу и в поисковый образ документа (ПОД): ведь в документе обсуждается не любовь, а какая-то научная проблема. Даже если цитата находится в «сильной позиции», являясь заголовком, а в тексте многажды подтверждена важность ее в целом и по частям, к ней неприменимы вопросы, которые можно поставить к описанию любой стандартной физической ситуации. Так, к статье двух авторов с заглавием «*Лодку унесло ветром»* (Мустайоки, Копотев 2005), например, не нужно задавать вопрос *Где и когда это случилось?* Нелепым будет и вопрос «из жизни», прагматический (*А был ли кто-то или что-то в лодке*?) или другие уточняющие эту ситуацию вопросы. Между тем такое смешение уровней вовсе не курьез в обсуждениях лингвистами семантики лексем и высказываний (см. пример в разделе 12), а в жизненных беседах и даже в передачах СМИ очень часто происходит зацикливание на семантике приводимого примера, что почти всегда приводит к уходу от основной нити (логики) разговора.

Вернемся к специфическому «жанру» заголовков (ЗГЛ). Одиночные лексемы или термины в заголовке всегда неполны: *Эйфория. Возвра-*

щение. *Изгнание* (Кого? Куда? За что?). И вывески на магазинах и прочих учреждениях заведомо неполны: «*У Николая*», «*Для ветеранов*». Даже если ЗГЛ имеет вид синтаксически полного предложения, чаще всего отсутствует явно выраженная связь с самим текстом (*Многое еще впереди*; *Тише едешь – дальше будешь*; *Не в свои сани не садись*; *На палубу вышел, а палубы нет*). Этот вид компрессии – обобщение – вряд ли скоро будет доступен для автоматической обработки, хотя какие-то выводы Системы уже умеют делать. Системы машинного перевода (МП) делают порой катастрофические ошибки, если перевод текста начинают с перевода ЗГЛ. Так, ЗГЛ *General programs* одна система МП (с англ. яз.) перевела как *Генерал программирует*. Другая система МП (с франц. яз.) предложила для подзаголовка *Alimentation de Pericles-Michelet* несколько интерпретаций, требующих такого перевода: *Питание/пища Периклеса-Мишле* или *Метод (технического) питания по Периклесу-Мишле* и др., тогда как речь шла об установке передачи электроэнергии из пункта Периклес в пункт Мишле. Неполнота в заголовке к естественному тексту – неизбежное и вполне законное свойство жанра ЗГЛ. Только аппарат семантического анализа (СемАн) может эксплицировать ее, отразив в промежуточном СемП то или те смысловые отношения, которыми ЗГЛ связано с какими-то элементами текста. Но для этого СемАн должен поставить соответствующий вопрос в виде формулы с пустым местом, которая участвует в дальнейшем анализе текста на равных правах с другими локально не заполненными валентностями.

Также вполне законны заголовки книг, выраженные только предложной группой: *В лесах, На горах*. В реальном газетном ЗГЛ *«Россия в обвале, Запад в ужасе, кризис в разгаре, Ельцин в Барвихе*» все четыре неполных предложения интерпретируются как корректные смысловые формулы: СУБ(*Россия,обвал*), ПАЦИЕНс(*Запад,ужас*), СТАДИЯ(*разгар,кризис*), ЛОК(*Барвиха,Ельцин*). А жанр ЗГЛ не требует уточнения по времени, как того требовали бы те же высказывания в составе текста. Как всякий заголовок, они требуют лишь обоснования связи СемП (ЗГЛ) и СемП тела текста. Наблюдения показывают, что в газетах утверждается тенденция давать достаточно коротким текстам о событиях полные ЗГЛ, напоминающие аннотации (*Маньяк убил старушку за пенсию*).

Но самое большое количество неполных высказываний, да еще с нарушениями грамматики, поставляет разговорная речь, особенно отличаются в этом политические деятели: *Дать от микрофонов по порядку ведения*! (спикеры в Думе) или *Мы обменялись и здесь. Участникам надо добавить – и будем добавлять. Прибавим во всех отношениях*. (Черномырдин, 9 мая 1997 г.). Это такая ситуативная неполнота, когда всем ясно («по жизни», а не из текста), что будет добав-

ляться и кому, но Система объявит подобные высказывания эллиптичными и неграмотными.

5. Неполнота относительно состава словаря

Первый обязательный компонент Системы АПТ, *относительно* которого можно диагностировать неполноту высказывания и текста, – это Словарь.

Поиск в словаре всех выделенных графематическим анализом текстовых единиц обычно совмещен с морфологическим анализом словоформ. О любой неудаче Система должна сигнализировать и предлагать какой-то выход. Если текстовой единицы не оказалось в словаре, то диагнозом может быть «неполнота самого словаря относительно текста», и эти случаи (пополнение словаря или коррекция морфологии) рассматриваются отдельным процессором, не интересующим сейчас нас. Важно то, что любой неопознанный лингвистический объект (НЛО) или просто новое слово прерывает анализ или оставляет пробел (знак вопроса/неопределенности) в семантической структуре. НЛО может оказаться словом с ошибкой, или содержащим в составе одного слова буквы разных алфавитов, что возможно как по причине путаницы в регистрах, так и специально, как языковая игра (Аkademия, Яndex в разных вариациях). В том и другом случае Система примет его за ошибку и сделает вывод о смысловой неполноте предложения или фрагмента. Она может использовать (для очень частых «ошибок») какой-то корректирующий модуль и продолжить анализ. Плохо, если вывод сделан неправильно и анализ пойдет дальше по ложному пути (например, *Коперник* превратится в *соперник*а, либо наоборот).

А что делать лингвистической Системе с таким специфическим корпусом, как тексты реклам? Ведь они содержат порой много новых слов. *Запепсуйте мегахит*! – такой плакат был одно время развешан по всей территории МГУ. Содержание его можно понять только из картины, где молодые люди с идиотическим выражением лица опрокидывают в глотки бутылки с Пепси, одновременно выписывая кренделя ногами в коньках. Все выражение может быть отнесено к категории НЛО, так как у него нет ни контекста, ни валентностей. Лингвистический анализ может что-то прояснить. Так, предложение будет расценено Системой как содержащее лексическую неполноту, так как этих слов нет ни в одном словаре. Морфологию и синтаксис Система может приблизительно «вычислить». Семантика поймет, что это призыв к чему-то, что дано на рисунке, и даже предположить, что призыв имеет отношение к понятию *Пепси* (на бутылке есть такое слово), что призывают к чему-то современному (*хит*), да еще в модальности «очень» (*мега*). Проводить семантический анализ (СемАн) реклам типа «*Азбука*

вкуса»; «*Займитесь стейком с профессионалом*!» неэффективно, пусть этим занимается другая наука (возможно, социология), но вопрос, что делать с новомодными словами, остается: надо ли вводить в основной словарь все новые лексемы, которые появляются ежедневно в средствах СМИ, или дать им укорениться?

Конечно, неполнота высказывания в тексте объясняется часто неполнотой словаря, и хотя это может быть объявлено «временной трудностью», она влияет на общую оценку текста по признаку СНТ. Большое количество НЛО в тексте снижает оценку полноты и связности всего текста. Выходом было бы не столько расширение объема словаря, в том числе за счет создания «специальных» словников, сколько увеличение силы семантического анализа таких «поверхностных» явлений, как НЛО.

Связанная с этим, но более сложная проблема – определение СНТ относительно **содержания** словарной информации (см. Раздел 8). Но прежде коротко остановимся на неполноте синтаксического уровня.

6. Эллипсис как неполнота относительно синтаксической структуры

Термином «эллипсис» мы обозначаем явление только синтаксического уровня, а именно, нарушение синтаксической структуры (или СинП) предложения из-за опущения одного или ряда слов (Леонтьева 1965, и др.), являющихся дублями слов ближайшего контекста. Неполное СинП приводит и к неполноте СемП, но лишь на коротком отрезке текста. Явление СН намного шире, чем только синтаксический эллипсис, и отождествлять их нежелательно. Эллипсис легко диагностируется, опущенные слова достаточно просто восстанавливаются из соседнего, как правило, симметричного с ним, предложения (чаще непосредственно слева). Примеры: *Твоя **дача построена** у дороги, а моя – в глубине леса*. Эталоном для сравнения считается типовая структура простого предложения, которая включает сказуемое, сильные синтаксические актанты, в том числе подлежащее, и (факультативно) сирконстанты. Учитывая поправки на жанр, требования к правильной (скорее, приемлемой) синтаксической структуре могут несколько варьироваться. Но можно и не обращаться к нормативному СинП, а сравнивать структуру эллиптичного предложения с симметричным ему, одновременно восстанавливая лексемы (в данном случае ***дача построена***). Это грамотные эллиптичные предложения. Если же разорванная синтаксическая структура не допускает восстановления опущенных лексем, предложение считается неграмотным – именно такой сигнал должна дать Система.

На синтаксическом уровне наиболее часты случаи неполноты в именной группе за счет опущения главного имени при наличии его оп-

ределения: *На **красный** ездит тот, у кого много **зеленых**. На **Первом** за освещением выборов будет присматривать телевизионщик Андрей*. Еще пример (дуэт из оперы):

Я красную просил сорвать. – Какую это? я не знаю. – Одну из тех просил я дать… – Какую? я не понимаю. Верни мне ту, что я дала, и я сорву тебе другую…

Один из выходов при анализе таких эллиптичных фраз – вводить символ отсутствующего имени существительного для полноты структуры. Второй случай – неполный состав актантов: главное слово остается без части зависимых. Здесь локальная неполнота скорее всего прояснится при обращении к другим фрагментам высказывания. Третий случай – когда отсутствует существительное при наличии управляющего им предлога: *Сколько раз мы молчали по-разному*: *Но не **против**, конечно, а **за**…* (Галич). Он приемлем как художественный прием (ср. *Он взял меня за и втянул в…* – из письма человека, которого заставили участвовать в какой-то конференции), но получит низкую оценку, если встретится в деловом тексте.

Если производные слова одного лексического гнезда законно наследуют значения полей МУ и ВАЛ (см. ниже) от главного полного предиката, то иногда наследование идет еще дальше: не от «родных» слов того же лексического гнезда, а от родственников, причем с противоположным значением. Пример: *молчать **за*** или ***против*** – по аналогии с *высказываться*, *говорить*, *голосовать **за*** или ***против***. Ср. еще: *Уже 2 месяца я молчу **про** нашу бурно текущую жизнь* (из письма Ф.). Или: ***О чем** молчал памятник Гоголю* (ЗГЛ).

Проблема синтаксического эллипсиса достаточно освещена в теоретической и прикладной лингвистике. Нас в данном случае интересует, как должна отображаться (какой след оставляет) такая неполнота в СемП предложения и как с ней бороться при установке на полный анализ целого текста. В основном это случаи совмещения эллипсиса с другими видами неполноты (вызываемая им неоднозначность и др.).

7. Компрессия и неполнота в сочинительных конструкциях

Еще один результат законного сжатия являют собой однородные члены предложения: *Министры {иностранных дел и обороны} {Игорь Иванов и Игорь Сергеев} приступили к обсуждению вопросов, связанных с войной за рынки сбыта.*

Сочинительные конструкции всегда являются результатом сжатия, полный вид конструкции в данном случае был бы таким: *Министр иностранных дел Игорь Иванов и министр обороны Игорь Сергеев*…Он был бы проанализирован Системой корректно. А сжатая фраза требует некоторого, хоть и простого, вывода, опирающегося на

знание названий возможных должностей высшего ранга (что реально в Системе), а также на учет симметрии (2 и 2) членов однородных конструкций. Такое же правило может проработать на фразах типа *Петя, Игорь и Саня дали Маше, Оле и Тане яблоко, грушу и сливу соответственно*. Слово-оператор *соответственно* подсказывает алгоритм обработки таких конструкций при анализе: *Петя дал Маше яблоко* и т.д.; фраза может быть признана синтаксически корректной.

Разрывные союзы и парные предлоги вводят конструкции, содержащие подобный же вид законного сжатия: *В ночь с 12-го на 13-е марта произошло это покушение.* Синтаксический анализ работающих Систем должен правильно разбирать такие конструкции и не засчитывать им лишнюю СН.

Но есть и некорректные способы сжатия в однородные группы, например, *Уступайте места пассажирам (старшего возраста, с детьми и инвалидам)*. Развертка такой конструкции при переходе к СемП дает право поставить вопрос: *пассажиры старшего возраста*, они же *с детьми*, и они же *инвалиды*? Или это три разных актанта? В подобных случаях фраза получит низкую оценку по синтаксису: в сочинительной группе нет синтаксического согласования.

8. Зависимость оценки неполноты от информации в словаре

Задаваемая в словаре семантика лексемы полнее всего проявляется в валентностях, приписываемых этой лексеме. Упростим картину классификации валентностей, которая не слишком проста теоретически: сложность ее прекрасно продемонстрирована в (Mel’čuk 2003). Будем в прикладной Системе различать всего два уровня валентной структуры. Один, вводящий синтаксические модели управления с морфологическим способом их выражения, так и обозначим: МУ. Другой уровень – это предсказываемые значением лексемы имена семантических связей, или семантические валентности; обозначим их СемВал, или просто ВАЛ.

Главными с точки зрения семантики лексем являются валентности смыслового уровня, в терминах Ч. Филлмора их можно считать семантическими падежами. Анализ многих лингвистических работ показывает, что часто между количеством МУ-значений и ВАЛ-значений ставится знак равенства, т.е. достаточно над каждым способом сильной связи синтаксического управления поставить его семантическое имя – и обеспечено полное описание семантических связей лексемы. Однако слишком многие примеры не соответствуют этой простой модели.

В нашей модели разрывается жесткая связь полей МУ и ВАЛ: между количеством единиц в полях «ВАЛ» и «МУ» семантического словаря нет прямого соответствия. Синтаксические модели управления

привлекаются как способы реализации СемВал только в том случае, если смысловая связь выражается каким-то фиксированным морфолого-синтаксическим способом (МУ). Многие поля ВАЛ часто и не сопровождаются заданием МУ, а остаются просто семантическими предсказаниями, которые в словаре уточняют описание определенного значения лексемы, продолжая ряд приписанных лексеме таксономических, иерархических и других семантических характеристик (обозначим их коротко СХ). В тексте эти предсказанные полем ВАЛ связи могут реализоваться самыми разными способами. Некоторые лексемы имеют записи в поле ВАЛ, для которых МУ вообще не выражаются: *дезинформация*, *СМИ* (Леонтьева 2006а и др.).

Процедура семантического анализа с оценкой типа неполноты каждого отдельного высказывания *относительно* информации словаря на первый взгляд проста. Она начинается сравнением фразового контекста всех его лексем со словарем. Несовпадение синтаксического контекста во фразе с описаниями МУ в словаре можно квалифицировать как синтаксическую неполноту (она же часто дает при интерпретации и смысловую неполноту). А несовпадение семантических характеристик слов во фразе с СХ, предсказанными в словаре (поле ВАЛ), квалифицируется как семантическая неполнота. Простота этого принципа оценки – только кажущаяся. Даже если настроиться на самый прозрачный и монотонный жанр – деловых документов, то и без подсчетов ясно, что полная МУ чаще не реализуется в тексте, чем реализуется. В словаре полный состав МУ как формальной репрезентации актантов задается для личных форм глагола. Заведомо меньше элементов МУ бывает при неличных формах глагола – у возвратной формы, инфинитива, причастия, а также у отглагольного существительного, и невыраженность отдельных актантов при них вполне естественно объявить законным (грамотным) видом не только синтаксической, но и смысловой неполноты. Но ведь даже если все объявленные в словаре грамматические формы МУ нашлись в тексте, это еще не значит, что они и есть актанты. Если основываться только на совпадении грамматических форм, мы найдем много ложных «актантов» при отсутствии необходимых. Ср. *Его обвинили в сговоре* ***с*** *отмыванием денег*. *План Лаборатории МП* ***с*** *французского языка* ***на*** *1989 год*. При совпадении требований к МУ здесь явная смысловая неполнота из-за несовпадения СХ, найдены «ложные» актанты.

Первичным в словарном описании должно быть задание семантики связей. Вообще-то семантику связей задает Грамматика СемО, но в словаре в поле ВАЛ могут вноситься индивидуальные для лексемы дополнения или корректирующие правила. От значений ВАЛ зависит и набор элементов МУ, и степень обязательности их выражения при разных словоформах данного лексического гнезда, и другие условия их

использования (совместимость/несовместимость разных способов и др.). Поэтому мы и ведем описание семантических связей лексем «сверху вниз».

Словарь не ограничивается заданием только тех связей, которые требуют синтаксически сильного управления, а дает описание сильных элементов той СИТуации, которую описывает данная лексема в данном значении (см. раздел 10).

9. Оценка неполноты высказывания относительно Грамматики СемО

Список элементарных двуместных отношений с фиксированными требованиями к заполнению мест составляет Грамматику семантических отношений (СемО). Грамматика вводит ограничения на СХ членов отношений, например, СемО ПРИЧИНА(А,В) будет признано полным и правильным, если на места А и В попадают лексемы с нужными СХ, в данном случае оба СХ – «Действия» или «Ситуации». Обычно требования к СХ формулируются или корректируются в поле ВАЛ. Так, слова класса *Институт*, *Школа*, *Завод* и другие с СХ=«Учреждение» имеют ВАЛ=СПЕЦиализация(А,«Учр»), где на месте семантически зависимой лексемы А ожидается СХ=Область деятельности («О_Деят»), например, *Институт физики*, или СХ=«Действие», например, *Институт повышения квалификации*… Это соответствует Грамматике СемО, в отличие от вывески на остановке «*Институт зерна*» или от имени учреждения «*Институт картофеля*», которые будут интерпретированы семантическим анализом как неполные: СХ(*зерно* или *картофель*)=«Вещ-во» (вещество). В этих случаях должны работать послабления жанра вывески, ЗГЛ и др. Эти названия будут оценены как имеющие неполноту, но законную, что не мешает «высветить», вербализовать те вопросы, которые остаются в промежуточном СемП (*Какова специализация Института*, *Что он изучает*? и т.п.). Ответы позволят восстановить полный вид примерно так: «Институт исследования проблем, связанных с зерном», или даже «связанных с выращиванием и другими действиями с зерном…».

Приведем другие примеры нарушения грамматики СемО: *Канал должен контролировать предоставление эфира. Партия может подать в суд на канал. Это известно от источника. Источник в Кремле сознался, что…* Слова *Канал*, *Партия*, *Источник* и т.п. принадлежат (это зафиксировано в словарной БД) к категории семантически зависимых слов: они представляют не самостоятельный объект, а лишь какой-то аспект, параметр и т.д. самостоятельного объекта, занимая первое место в составе СемО: ЧАСТЬ(*Канал*,Х), ИСТОЧ(*Источник*,Х) и т.д.

Имена СемО – это основа используемого нами (пусть временного, в отсутствие общепринятого) метаязыка. Хотя эти двуместные отношения выражены с помощью лексем естественного языка, они не равны словам ЕЯ, у них есть свойства, приближающие их к формальным языкам (эта тема развивается подробнее в других работах автора). При создании списка отношений соблюдался принцип «умеренной универсальности», я с удовольствием заимствую этот термин из работ А. Мустайоки (2006, 37). Мы вводим минимальную формализацию, в основном касающуюся лишь ограниченного синтаксиса записи. Полнозначные лексемы с высоким словарным весом: 4 или 5 по 5-тибалльной шкале (Леонтьева 2006а) представлены словами ЕЯ.

Смысловое отношение с заполненными по тексту членами отношения – это только минимальная единица текстовой семантической структуры. Следующая, средняя, единица семантического метаязыка – текстовая Ситуация (СИТ). Она может дополнить список словарных смысловых валентностей данной лексемы.

10. Оценка неполноты высказывания относительно структуры Ситуации

СИТ – это **лингвистическая** единица, репрезентирующая состав и содержание простого предложения. Несколько СИТ, отражающих относительно законченный фрагмент содержания текста, можно объединить, подчинив их единице следующего уровня, нетерминальному символу «Высказывание» – ВЫСК (им может быть и абзац, и сложное предложение, состоящее из нескольких простых). В достаточно простом случае все Высказывания подчинены высшей единице – ТЕКСТ – непосредственно или через цепочку однородных с ними единиц. В сложных текстах может быть гораздо больше промежуточных единиц (нетерминальных символов, или НТС), например, Главы, Разделы, Части и т.д.; сноски, примечания и пр. должны образовать особые единицы в связке с той единицей, от которой есть ссылка, и так далее. Такая композиционная бухгалтерия вполне достижима в прикладных системах. По сути дела этим фиксируется макросинтаксис текста, и он повторяет систему зависимостей и составляющих синтаксического уровня, только в масштабе текста как целостной единицы.

Но в данной работе нам достаточно рассмотреть «среднюю» единицу СИТ, так как ее структура воспроизводится во всех следующих более сложных. Итак, полная типовая структура СИТ и правила построения ситуативного представления (СитП) для простых двусоставных предложений (которых в тексте большинство, учитывая и результаты сегментации сложных) таковы. Нетерминальный символ СИТ подчиняет лексему, выбранную как лексическое ядро (ЛЯ) ситуации. В

стандартном случае им становится главный предикат. Если у него максимальный информационный вес, такое СитП является устойчивым, его ЛЯ переходит в структуру со своими смысловыми валентностями, частично заполненными на предшествующем уровне анализа (СемАн1). Единица СИТ тоже имеет словарную статью со своим набором валентностей, это перечень отношений, которые могут характеризовать любую ситуацию: ВРЕМЯ(?,СИТ), ЛОК(?,СИТ), УСЛ(?,СИТ), ПРИЧ(?,СИТ), УТОЧН(?,СИТ) и другие значения синтаксических, в основном сирконстантных и «слабых», связей. Состав сильных валентностей единицы СИТ существенно определяется тем, какую СХ получила сама СИТ – учитывается СХ ее ядра и предметная область, жанр и иногда задача: для физических действий в модальности РЕАЛ это УТОЧНение(,СИТ) по месту и времени, для абстрактных СИТ – это УСЛовие(,СИТ) и т.д.

Логическая сумма наборов СемО, являющихся валентностями единиц СИТ и ЛЯ, образует гипотезы относительно присоединения к этим символам других объектов, в том числе оторванных членов предложения, что позволяет построить максимально полную и связную структуру.

Наибольший вес в словаре имеют лексемы, обозначающие активные действия (*строить*, *разрушать*, *воевать* и т.п.). У многих предикатов вес зависит от семантики актантов (чаще это смысловой объект или содержание), что фиксируется в нашем словаре. Важны связи модальности и оценки в СитП. Модальность «РЕАЛ» (реальная Ситуация) повышает вес, а модальности, отрицающие само действие, снижают вес до нуля.

Если вес главного предиката оказался ниже веса зависимого от него предиката, СитП неустойчиво и требует перестройки иерархий, установленных в СинП и унаследованных структурой СемП1. На роль ЛЯ выйдет зависимое; главное же слово СинП (и СемП1) станет в структуре СИТ зависимым, а его собственная семантическая характеристика будет именем связывающего их СемО. Так, *подготовка к сбору урожая* перейдет в СитП как *сбор урожая* в стадии *подготовки*, поскольку ВЕС слова *подготовка* ниже ВЕСа узла *сбор урожая*. Этот последний станет лексическим ядром СитП, а синтаксически главное слово – семантически зависимым; в данном случае именем связи будет собственная характеристика слова *подготовка*:

СТАДия (*подготовка, сбор_урожая*).

Для синтаксически неполных предложений будет построено заведомо неполное СитП. Предикат односоставного предложения (например, *Война.*) займет место ЛЯ, а валентности узлов СИТ и ЛЯ останутся пока незаполненными. Многие лексемы и выражения с очевид-

ной семантикой займут соответствующее их основной характеристике место в структуре СИТ, например, односоставная фраза *Осень 1993*. встанет на первое место СемО ВРЕМЯ: **ВРЕМЯ**(Осень_1993,**?СИТ**), т.е. на данном отрезке текста остается неясным, о какой ситуации пойдет речь в тексте. Даже в случае односоставных предложений с «предметными» лексемами (*Река. Нефть. Рыба.*) им отведена роль пока неясного АКТАНТа неизвестной СИТ. Выяснение того, к каким ситуациям присоединить такие изолированные узлы, будет отнесено на следующие этапы. Сама фиксация семантически необходимых, но отсутствующих на данном отрезке участников Ситуации создает формальный стимул для перехода к анализу целого связного текста.

11. Оценка неполноты высказывания относительно Знаний

Чтобы оценивать полноту-неполноту текста относительно Знаний, нужно иметь непустой массив структур, представляющих знания (назовем их ЗнП, по аналогии с СемП и другими лингвистическими структурами). Пока можно говорить лишь о неполноте или даже отсутствии тех записей знаний, с которыми можно сравнивать содержание анализируемых текстов. Рассмотрим два несовместимых взгляда на то, какой вид может принять структура Знаний.

Специалистам по искусственному интеллекту (ИИ), в частности, занимающимся системами извлечения знаний из текста, важно построить на выходе базу данных (БД), а для текста – такую семантическую сеть или *концептуальную* структуру, в которой остались бы только интересные им (актуальные для них) единицы: важные объекты и отношения. Это *концептоцентрический подход* (concept based approach), не допускающий в своих структурах ничего «лишнего», не заказанного заранее; естественно, в нем и нет средств для отображения членения текста на предложения и другой «мелкой» лингвистической информации. Для лингвистов же цель построения правильного СемП состоит в том, чтобы сохранить в нем все тонкости и нюансы, которые передала языковая материя в тексте. Лингвистический взгляд на задачу остается неизбежно *синтаксически ориентированным* (syntax driven, syntax oriented), так как сохранение всей текстовой информации при переводе в СемП требует сохранять и авторскую упаковку этой материи в отдельные предложения, абзацы и т.д.

Можно ли вообще согласовать эти две позиции, т.е. строить такую семантическую структуру текста, которую лингвисты и когнитологи могли бы признать приемлемой или даже назвать ***правильной***? На этот вопрос, а также на следующий «Как конкретно ее строить?» ответ должна дать – в рамках вербальной системы – **лингвистика**. Именно она как более тонкая организация может подсказать, как обойти – при

решении практических задач – многие ее тонкости. Иначе они будут сметены (да простят меня представители ИИ-наук!) «грубой силой». Информатика работает с парадигматическими отношениями, но не умеет работать с синтагматическими. Лингвистика не выходит на уровень построения крупных единиц (и узлов, и отношений), которые могли бы вывести ее в сферу межтекстовых категорий и связать с разными БД, описывающими разные области специальных знаний. Конфликт методов лингвистики и информатики, продолжающийся уже 40 лет, может и должна преодолеть *прикладная лингвистика*. Перед ней стоит задача (или, что одно и то же, она стоит перед задачей) – НЕ нарушать лингвистические законы построения ее главного объекта исследования – естественного текста (ЕТ), а скорее опираясь на них, уметь представить во внешнюю среду основное содержание ЕТ. Говоря более высокими терминами, важно сохранить **экологию** ЕТ во всех формальных манипуляциях с ним. Это означает: Не дать увидеть в тексте то, чего в нем нет, не вынести в главное содержание текста то, что сказано «между делом» или по оплошности, короче, не исказить смысл текста, но уметь извлечь из него ЗНАНИЕ, которое можно положить в «копилку человеческих знаний». Построить такой анализ целого текста – сложная задача, но она по силам сообществу прикладных лингвистов, уже работающих с текстовыми корпусами.

Предложенное в (Леонтьева 2006б) предварительное (полуформальное) описание текстовых содержательных единиц СИТ, СОБ (событие) и ТФ (текстовый факт) имеет целью наметить контуры структуры ЗнП, общей для любых текстов на естественном языке (ЕЯ). Пока реально говорить о создании СитП для связных фрагментов текста (например, абзацев). В пределах этих СитП работает механизм взаимного насыщения неполных формул и «гашения» имеющих малый вес валентностей. Сжатые таким образом СитП соседних абзацев могут сравниваться друг с другом формально, поскольку они строятся одними и теми же алгоритмами. Так будет моделироваться процесс накопления Знаний самого текста и оцениваться степень неполноты отдельных его участков (или незнаний, вопросов) **относительно** уже накопленных фрагментов Знания. Этот алгоритм ляжет в основу сравнения текстового содержания с будущими структурами представительного ЗнП.

12. Неполнота, вызываемая смешением уровней описания

Смешение уровней в лингвистических рассуждениях – достаточно сложная проблема, чтобы касаться ее «между делом», но все же один пример я приведу. Рассмотрим следующий фрагмент из книги (Добрушина и др., 2001, 130):

«[1] *он берет книгу/бутылку/шляпу/*…

S агенс, представляющий человека; A – объект, который можно перемещать, которым можно манипулировать; характер стабилизации выводится из естественного для S способа стабилизировать A – «пальцами» <…>

[2] *он берет трубку*

Толкование данного примера очень близко толкованию [1], с той лишь разницей, что *трубка*, благодаря своим специфическим характеристикам, сохраняет тесное отношение со своей начальной позицией (телефонный аппарат), являющейся референтной позицией A (Q’’3): завладение S термом A происходит лишь на время телефонного разговора.»

Человек, скорее всего, поймет этот пассаж правильно, но «умный» автомат найдет много поводов для «возражений» или сигналов ошибок. Возьмем только последние 4 строчки. Слово *трубка*, выделенное курсивом именно как **слово** ЕЯ, оказалось связанным с предметом из **жизни** (телефонный аппарат), который, в свою очередь, есть **референтная позиция A**. Здесь сразу три ошибки: **A** как предмет, **A** как терм и **A** как возможное имя позиции, при том, что выражение *референтная позиция* само является некоторым или концептом, или понятием метаязыка. А на интерпретации выражения «*завладение S термом A*» автомат просто сломается: *завладение S* как человека? Так это физическое действие?! Тогда правильнее *человеком S*. Но причем тут *терм A*? Это ведь элемент метаязыка, им нельзя завладеть, да еще только «на время телефонного разговора», т.е. в жизненной ситуации. Все подобные «уровневые» противоречия требуют явного выражения в виде неполных формул СемП, или вопросов к автору.

Этот мысленный эксперимент семантического анализа пока нельзя реализовать, а если удастся, то он поставит на этом же фрагменте еще больше вопросов. Пока же он позволяет думать о том, какого рода инструментарий надо готовить, даже если не стремиться к предельной задаче СемАн текста, а использовать полуформальный метаязык и модель СемАн для контроля собственных формулировок. Так, даже для анализа этого короткого обрывка текста нужны как минимум три типа «словарей»: словари слов/лексем ЕЯ, словари-тезаурусы концептов, словарь и грамматика метаязыка. Что касается «предметов и явлений действительности», на которые часто ссылаются лингвистические работы, то они могут быть учтены в интеллектуальной системе только в виде имен соответствующих им концептов, включенных в концептуальные словари-тезаурусы либо в БД и базы знаний.

Предельные случаи смешения уровней проявляются в микротекстах лозунгов и объявлений, когда соседствуют вербальные и невербальные способы: Например, в Киеве над дорогой была такая растяжка: Знак кирпича (как невербальный способ, запрещающий проезд) продолжен словами «*и по тротуару*». Или *Я* + (знак сердца, пронзенного стрелой, как символ любви) + *Москву*. При анализе такого рода высказываний, если они попадают в корпус, Система должна ставить вопросы, отсылать к соответствующим ПО и т.д.

13. Неполнота логическая

Кроме перечисленных видов неполноты, можно назвать неполноту логическую, когда одновременно делаются два или больше противоречащих друг другу утверждения (*На палубу вышел, а палубы нет*) или части утверждений: *Ночи и дни целыми ночами они думали, как вырваться из плена* (Рыбкин); *Наконец в этом году у нас произошло снижение повышения роста падения* (Черномырдин); *Милиции больше, чем людей* (разг. на улице); *Не удалось наконец-то нам в данный момент остановить преступность* (Гл. прокурор, речь по радио). В последнем примере несколько довольно тонких нарушений логики, демонстрирующих несовместимость четко выраженного значения длительного действия (*Не удалось… остановить преступность*) со словами (*наконец-то* и *в данный момент*), характеризующими моментальные действия. (Попытка сформулировать правило несовместимости этих значений сразу нарушает постулат о едином уровне рассуждений, см. выше раздел 12.)

Конечно, Системы искусственного интеллекта (ИИ) еще не доросли до того, чтобы даже ставить задачу хотя бы выявления подобных смысловых конфликтов. Но некоторые формальные противоречия можно устранять обращением к фрагментам ЗнП, как в услышанном мною примере *Он был ректором Петербургской, Петроградской, а затем и Ленинградской консерватории*. Здесь нельзя сделать вывод, что «некто **он** часто менял место работы», так как БД «Географии» даст сведения о том, что все три имени называют один объект и локализация объекта не меняется. Более того, референт лексемы *он* тоже можно установить из обращения к имеющимся спискам важных должностей и связанных с ними имен (по годам). (Видимо, это композитор Мясковский).

Много противоречивых утверждений встречается в жанре сказок, например: *Принц Набюссан обладал несметными сокровищами. Но его обкрадывали и обманывали. Половина богатств принадлежала лично ему, а бОльшая часть – его администраторам*. Итак, богатства нельзя пересчитать, в то же время упомянута половина того, что неисчислимо,

и остается еще бОльшая часть от половины. Семантический анализ может поставить вопросы (половина чего?, бОльшая часть чего?), а разъяснение должна дать соответствующая предметная область (ПО = арифметика), если Система умеет общаться с ПО. Остается повторить, что Системы ИИ пока не обладают интеллектом, нужным для логических операций с вербальным материалом. Конечно, сначала должна проработать поправка на жанр, и для анализа сказок не нужно выходить ни в какие точные ПО. Но такого рода семантические казусы встречаются и в деловых текстах, и их нужно уметь хотя бы обнаруживать.

14. О неграмотности

Если в текстовых корпусах классической литературы лингвистами наводится полный грамматический порядок, то в массивах современной «деловой» документации, а тем более в текстовых корпусах уличных и настенных объявлений, в публикациях и устных разговорах и даже в передачах СМИ мы вынуждены сталкиваться с большим количеством нарушений грамматических и стилистических норм. Это приводит к тому, что оценка «неполноты» будет для них завышена. Самыми очевидными нарушениями норм является неправильное употребление падежей и предлогов.

М.Я. Гловинская в серьезном исследовании тенденций ЕЯ в XX веке (2000, 251) говорит об «экспансии» предлога *по* (см. *Долги России по газу* вместо *за газ* и другие ее примеры), а также о чрезвычайном натиске на глаголы предлога *о* (*Об этом отмечается…* вместо *это*, и мн.др.). К названному дефекту добавим еще и фактор избыточности поверхностного выражения мысли: *...то, что…* или *...о том, что…* вместо простого союза *что*: *Не удивляйтесь о том, что в Думе появились собаки* (Селезнев, которому сообщили, *что в Думе заложена бомба*). *Я понимаю о том, что Явлинскому трудно удержать свой электорат…* (радио). *Мы имеем в виду о том, что… Мы удивляемся о том, что…* вместо *удивляться тому, что* (так указано в словаре). Чаще всего это неграмотный перенос модели управления (МУ) от лексем класса *говорить*. Избыточность часто наблюдается и в лексическом материале: *Имея соглашение о согласии президента, если он согласится* (радио).

Нарушенная МУ – будь то отсутствие нужной МУ или неправильная МУ – приводит к синтаксической неполноте, а как следствие, образуются «лишние» актанты. Но при семантической интерпретации СинП они перейдут в какое-то достаточно общее и приемлемое СемО: СФЕРА(А,В) или СОДЕРЖ(А,В) или даже УТОЧНение(А,В). В дальнейшем анализе слабые связи могут уточниться как значение сильной, но не заполнившейся (по разным причинам) валентности. Естественно,

что при генерации нового текста из полученного сжатого СемП можно и нужно соблюдать все правила грамматики.

Выводы

Способность какой-либо Системы при анализе разжимать текст, эксплицируя неполноту, и сжимать его семантическое представление (СемП), устраняя лингвистическое дублирование (которое имеется в любом тексте), была бы дополнительным свидетельством того, что мы имеем дело с анализом действительно семантическим, а не только стремящимся к нему. Именно СНТ держит здание текста как целого. Явления локальной неполноты интересны нам тем, что позволяют следить за семантическими процессами, которыми сопровождается восприятие и понимание текста. Кроме того, фиксация неполноты в промежуточных структурах служит и вполне практической цели – оценке качества создаваемого делового текста.

Мы рассматривали не столько информацию в текстовых корпусах, сколько отдельные примеры устной и деловой речи, которые могут также встретиться и в каком-либо корпусе. В задачи лингвиста входит также сбор самого корпуса трудных для понимания предложений. А если и когда он уже собран, такой корпус может служить тестовым материалом для проверки «прочности» синтаксической или семантической модели анализа.

Адрес электронной почты автора: leont.n@relcom.ru

Литература

Арутюнова, Н.Д. 1988. Типы языковых значений: Оценка. Событие. Факт. М.

Гиндин, С.И. 1982. К теории смыслосохраняющего сжатия текстов (ССТ). В кн. *Переработка текста методами инженерной лингвистики (Тезисы докладов)*, 65. Минск.

Гловинская, М.Я. 2000. Активные процессы в грамматике. В кн. *Русский язык конца XX столетия (1985–1995)*, 237–304. М.

Добрушина, Е.Р., Е.А Меллина., Д. Пайар. 2001, Русские приставки: многозначность и семантическое единство. В кн. *Исследования по семантике предлогов*. М.

Леонтьева, Н.Н. 1965. Анализ и синтез русских эллиптичных предложений. *НТИ. Сер. 2, № 11*. М. 41–46.

Леонтьева, Н.Н. 2006а. Автоматическое понимание текста: системы, модели, ресурсы. Учебное пособие. М.

Леонтьева, Н.Н. 2006б. Корпусная лингвистика: не только вширь, но и вглубь. В кн. *Труды международной конференции «Корпусная лингвистика – 2006»*, 234–241. СПб.

Мустайоки, Арто. 2006. *Теория функционального синтаксиса. От семантических структур к языковым средствам*. М.

Мустайоки, Арто & М.В. Копотев 2005. «Лодку унесло ветром»: условия и контексты употребления русской «стихийной» конструкции. *Russian Linguistics*, 1–38.
Mel'čuk Igor. 2003. Actants. В кн. *Meaning – Text Theory. First International Conference on Meaning – Text Theory*, 111–127. Paris.