Прикладные разработки

Н.Н.Леонтьева

Принципиальные особенности понятий «Тексты» и «Смыслы» в прикладных системах[1]

Обсуждаются трудности автоматического понимания текстов – разные виды неоднозначности и неполноты, структурной и смысловой. Предлагается фиксировать их в явном виде в новой структуре, названной Семантическим пространством текста; это своего рода расширение классического СемП И.А. Мельчука. Далее, если добавлены внетекстовые сведения, оно становится Информационным пространством этого текста. Неполнота в обеих структурах является движущим фактором всех дальнейших процедур «понимания», стремящихся к максимальной автоматизации.

Ключевые слова: автоматическое понимание текстов, смысловая неполнота, «мягкое» понимание, семантическое и информационное пространство текста, связность, сжатие содержания, множественность уровней понимания, адаптивная структура, индивидуальные смыслы.

0. Введение

Статья[2] посвящена трудностям автоматического понимания естественных текстов (АПТ). Все работы с текстами так или иначе охватываются двумя понятиями: ***Тексты*** и ***Смыслы***. Термин «**Смысл⇔Текст**» стал лозунгом и собственным именем известной лингвистической теории и модели (МСТ), описанной во многих публикациях, см. последнюю[3]. Но в слова и понятия *Текст* и *Смысл* каждая прикладная система вкладывает своё содержание (порой говорится лишь о морфологическом анализе словоформ). Поэтому в каждом новом описании работающих прикладных систем или их моделей необходимо уточнять: а. какие тексты предназначены для обработки и б. какой результат обработки назван «смыслом» (некая структура или другой, синтезированный текст, отличный от входного, или это восстановленная лемма).

Если речь идёт об анализе (от Текста к Смыслу) – какое множество текстов предполагается задать на входе работы Системы?

Произвольный ли массив текстов или множество фраз из одного текста? Каковы анализируемые тексты – научные или художественные? Если задан массив, то это однородный по тематике (по автору, по срезу времени и т.п.) корпус или политематический? и т.д. Если мы претендуем на анализ заданного произвольного массива текстов, например, массива текстов, появившихся за один день в СМИ и т.п., то понятие **Смысла** к ним плохо применимо: у массива, даже принадлежащего одному автору, НЕ может быть единого Смысла; в лучшем случае можно говорить о смысле одного текста или отдельных высказываний (абзацев и т.п.) из этого текста. Если же анализ претендует на извлечение Смысла из каждой фразы заданного Текста, то многие фразы не имеют смысла вне связи со своим окружением: мы можем получить только структуру, обычно синтаксическую или синтактико-семантическую, причём в задаваемой лингвистами нотации. Такая структура часто неполная, и назвать её Смыслом может только лингвист, но не обычный пользователь.

Если же идти в другом направлении (от Смысла к Тексту), то на каком (формальном) языке можно изобразить этот Смысл? До сих пор не создано ни одной убедительной интерлингвы для записи смысла произвольного текста или даже высказывания. Нет и единого языка представления Знаний (ЯПЗ) даже для ограниченного круга специальных текстов: каждая База данных по разным специальностям (химия, геология и т.д.) имеет свой специфический язык и свой способ фиксации знаний; это означает, что и анализ специальных текстов даст разнородные финальные структуры. Можно ли задать единую процедуру синтеза текстов ответов для всего множества таких разнородных и часто дефектных структур? Ответ: или нельзя, или только полу-эмпирически, что определяется информационной установкой и конкретной задачей.

А Читателю нужен и поиск своей информации в массивах текстов, и выявление основного содержания (Смысла) в одном тексте, и ответ на его запрос, желательно в виде естественного сжатого текста (аннотации, реферата, или даже перечня найденных в тексте Ситуаций и т.п.). Именно на такого читателя мы ориентируемся в предлагаемых моделях по возможности естественного, но максимально автоматизируемого «понимания» текстов (АПТ)[4].

1. Принципиальные ограничения «лингвистических» моделей

«Лингвистическими» мы назовём Системы, которые создаются в соответствии с моделью «**Смысл⇔Текст**» (МСТ). Сила реали-

зующихся «лингвистических» моделей состоит в следовании теории, как она определена в классических работах И.А. Мельчука и Ю.Д. Апресяна[5]. Из основных постулатов этой теории можно назвать: модульность, включая независимость анализа и синтеза, отделение грамматик от словарей, последовательный уровневый подход, формализованность, согласованность описаний грамматик и словарей и другие. Эти свойства желательно сохранять в любой прикладной системе, работающей с текстами.

Но вместо сакрального для МСТ слова *Смысл* как цели анализа и исходной структуры для синтеза лучше пользоваться его синонимом **Семантическое представление** (или **СемП**), тем более что в качестве Текста или его СемП обычно приводится пример одного предложения. Термин СемП более точный, так как отражает представление самого лингвиста об идее лингвистического смысла (ведь для бессмысленных фраз тоже строится СемП). Реализуя самый подробный анализ предложений, лингвистические модели сталкиваются со всеми стандартными «недостатками» естественных текстов. Достаточно назвать омонимию слов и словоформ, неполноту и неоднозначность (структурную и смысловую), метафоры, несвободные словосочетания, произвольный порядок слов в русских предложениях – они могут сочетаться в одном примере. Эти явления затрудняют применение программистских и математических формализмов и мешают построить однозначную структуру. Лингвистам постепенно удаётся справляться со многими из них, но устойчивым результат семантического анализа будет тогда, когда анализ выйдет за пределы отдельных фраз, в целый текст. Неоднозначность, перешедшая в СемП, может быть разрешена в составе Ситуаций, собранных и подтверждённых по всему тексту: ведь и анафорические, и разные логические связи пронизывают весь текст. При полном синтаксическом анализе их сборка начинается с задания так наз. «моделей управления» (таких моделей и систем мало[6]). Реализованные в пределах предложения Ситуации обычно неполны.

Не последним ограничением лингвистических систем являются достаточно жёсткие требования к входным текстам: это должны быть правильные предложения; с «плохими» фразами система плохо справляется. Аргументы типа «А так по-русски не говорят», объясняющие, почему Система НЕ перевела, НЕ обработала и т.д., не устраивают ни Авторов текстов, которые, при всех произволах стиля, желают получить хороший перевод своих текстов, ни обычных пользователей, которых мало интересует качество текста: им нужно получить полезную информацию.

2. О качестве текстов

Хорошими (нормальными) будем считать тексты, которые обладают регулярным синтаксисом и к которым можно с относительным успехом применить формальные грамматики; а незаконченные, с рваным синтаксисом, обрывки текста, случайные реплики и т.д. отнесём к трудным, плохим. Речевые тексты как правило плохие – они краткие, в них минимум контекста и поэтому они допускают слишком много интерпретаций.

Текст – это тоже речь, но только в письменном виде. Текст длиннее реплик диалога, он более организован, чем поток устной речи: у текста есть границы, композиция, текст привязан к определённому жанру, имеет фиксированное авторство, тексты доступны в машинном виде, и так далее. Текст оформлен грамматически и композиционно, он делится на кванты, регулируемые жанром: главы, разделы, абзацы, предложения и т.д.

Записанная речь даёт нам образец текстов, наиболее трудных для анализа и понимания. Приведу примеры таких отрезков, в основном это произнесённые по радио:

1. *В Москве этой ночью 11-13. Завтра от 21 до 23-х. Временами дождь. Местами сильный.*
2. *Мы делаем экскурсии по всей стране, кроме пятницы.*
3. *Нам за этот фильм дали по шапке.*
4. *Этот пейзаж отображён в двух картинах Писсаро, одной – Ренуара и одной – Моне.*
5. *Уборщицы прекрасного пола.*
6. *Мы приобрели 350 кг Ремизова* (= рукописи для музея).
7. *Россия **в** обвале, Запад **в** ужасе, кризис **в** разгаре, Ельцин **в** Барвихе* (заголовок в газете).

Неоднозначность и другие семантические трудности здесь очевидны. Приведу один короткий уличный диалог.

А: *Лежали и шли, а ночь полежали – и встали!*
В: *О ком Вы это?*
А: *Да о часах. Я их неделю назад купил, и гарантия есть.*

Ясно, что высказывание лица **А** непонятно без какого-либо разъясняющего контекста: скорее его можно понять применительно к кому-то живому. Узнав, что речь о часах, партнёр **В** поймёт слово *встали* как *остановились*. Это «плохой» текст, причём не последней трудностью является противоречие: невозможно одновременно лежать и идти (*Лежали и шли*). Здесь явное нарушение законов действительности, оно связано с отсутствием знания. Отнесение этого отрезка текста к предметной области «Часы» (как частный случай

базы данных или номенклатуры малых бытовых приборов) позволил бы уточнить значение многих слов: ***идти,*** о часах и о ряде малых приборов, означает **функционировать**. В семантических словарях лингвисты стараются указывать весь спектр значений слова, давая и правила перехода от одного значения к другому.

Гораздо труднее справиться с путаницей содержательных понятий в текстах, особенно когда идёт намеренное их искажение или даже прямая ложь. Эта «болезнь», более коварное свойство естественных текстов, получила в последние годы широкое развитие. Примеров в современном политическом дискурсе слишком много, чтобы надеяться разрешить их простыми средствами, как то обращением к Автору текста (например, «Автор высказывания ММ слово *враг* понимает как *соперник*, а если это же слово употребил НН, то скорее оно называет *сторонника справедливости*, и т.д.). Слова *либерал*, *демократ*, *патриот*, *иностранный агент*, *левые, правые, фашист* и многие другие можно считать утратившими свои исконные значения. Ясно, что ни теоретическая семантика, ни прикладная система НЕ справятся с такими социальными и текстовыми явлениями, как демагогия, прямое искажение фактов, просто ложь и даже ирония и юмор.

3. Некоторые классификации сложностей анализа

Не нужно думать, что такие загадки ставит нам только спонтанная речь. И в техническом тексте найдётся не один десяток отдельных фраз (предложения, заголовки, сноски, подписи под рисунками и т.п.), изолированный анализ которых не может привести к построению хорошей для них синтактико-семантической структуры. Автоматическое понимание заходит в тупик уже на этой стадии: для изолированной трудной фразы нельзя воссоздать **Ситуацию**, а её мы считаем главной единицей семантического анализа. Это внутритекстовый, **локальный тупик**.

Отсутствие компонента **Знаний**, общих и специальных, могущих дополнить не выраженные в тексте сведения – это более серьёзная проблема, а для лингвистики особенно важно отсутствие **аппарата сравнения** текстовых структур со структурами уже имеющихся профессиональных баз данных и знаний. Этот «**межтекстовый**» **тупик**, и он решается выходом за пределы текста. Только чтобы сравнить между собой содержание двух (любой пары) разнородных структур и включить сведения из одной в другую, нужен своего рода «машинный перевод» с языка знаний на язык лингвистического СемП и общезначимый язык-посредник[7], на который переводимо

любое знание – общее, текстовое и специальное. **Отсутствие стыковки** между тремя типами структур: жёсткими лингвистическими, более «мягкими» информационными и специальными, где каждая наука имеет свои единицы и свою логику, можно назвать **междисциплинарным тупиком**.

Проблема такого «машинного перевода», будь то естественное разноязычие или упомянутое выше внутриязыковое, не решена: она слишком сложна. Но её важно хотя бы поставить перед лингвистами – разработчиками прикладных систем.

Трудности правильного построения единиц разных уровней, отсутствие внешних источников знаний, иные логические и лексические противоречия на локальных участках текста постепенно можно преодолеть алгоритмическим путём. Но мы ищем решение, общее для систем, работающих с текстами, причём учитывая в первую очередь самый трудный материал. Именно трудный материал, изобилующий неполнотой, наталкивает на нестандартные решения там, где «чисто лингвистическая» модель заходит в тупик.

4. Другие структуры

Отсутствие зримых и давно предсказываемых нашей наукой результатов («смыслового» реферирования текстов, разумных ответов на вопросы к тексту, надёжного машинного перевода и других интеллектуальных продуктов), требует время от времени пересматривать установившиеся приёмы описания и способы включения необходимых компонентов понимания в процесс анализа. Я столкнулась с самыми реальными требованиями «немедленной помощи от «кибернетики» в учреждении с именем ВЦП, куда я перешла работать «со своими идеями» и необходимостью набирать новый коллектив. Коллектив почти сразу был вовлечен в технологическую цепочку «промышленных» заказов на перевод, и мы сами стали получать заказы на автоматический перевод текстов с французского языка на русский (отсюда название системы – ФРАП). Сначала мы настроили словари и систему на перевод текстов по металлургии, так как изучение литературы показало, что во Франции развита эта область и следует ожидать заказов на эту тематику. Получили резкую критику от вышестоящей организации (ГКНТ), сменили область текстов на микроэлектронику, нашли Заказчика и т.д. Жизнь заставляла либо переходить на почти-пословные и пословно-пооборотные модели перевода, по которым шли четыре другие лаборатории Всесоюзного центра переводов (ВЦП), либо искать совсем другой путь.

Учёт современных ожиданий от «умных систем» привёл нас к пониманию того, что без гибкой адаптации к меняющимся предметным областям работающую систему МП не построить. Молодой коллектив ФРАП был готов провести «теневой» эксперимент (наряду с выполнением основного плана) хоть и полного лингвистического анализа, но с «новым» семантическим компонентом, в структурах которого явно отражена Неполнота. Были несколько ослаблены требования к синтаксическому анализу и повышены требования к самому **началу** работы с реальным текстом: оно содержало много неполной, но важной смысловой информации[8].

Названный **другим** путь берёт за основу построенное синтаксическое или «синтактико-семантическое» представление (СинСемП). Но наше СемП отличалось от «классического» СемП (оно же **Смысл**) И.А. Мельчука тем, что в него входят и символы локальной смысловой неопределённости (неполноты) каждого анализируемого отрезка, как было предложено в работах 1967 и 1969 гг.[9] Впоследствии мы назвали его «первичным СемП», а сам анализ – «первичным семантическим анализом», см., например, статью Сокирко[10]. Виды неполноты дифференцировались: это и смысловые валентности значимых лексем, не заполнившиеся в составе изолированного отрезка текста, и валентности самого предложения, отражающие смысловые связи с предшествующим и/или следующим за ним предложением, и другие – из предшествующих этапов. Эту новую структуру мы назвали **Семантическим пространством** текста (сокр. СемПространство).

Формально все связи в нём описываются одинаково, формулой P(A,B), только часть формул, установленных или нет, будет неполной: P(-A,B), P(A,-B), P(-A,-B) или даже -P(A,-B) и т.д. Знак неопределённости говорит о том, что один из или оба члена отношения P(,) либо не нашлись при анализе, либо требуют уточнения значения (это касается и самого отношения P), либо не соответствуют правилам смысловой Грамматики. Существует градация видов неполноты, что можно изображать комбинацией знаков, например, ?P (-A?, B) и т.п. Единое отображение понятого Системой и непонятого позволяет выйти в Пространство целого текста.

Больше всего вопросов возникало на уровне первичного анализа, особенно это относится к специальным техническим статьям и политическим документам, в которых много аббревиатур и условных обозначений, расшифровка которых часто требует семантического анализа текста. Но и морфологический, и синтаксический, и все промежуточные между ними уровни порождают много вопросов, а это неполнота, апеллирующая к Семантике целого. Это и привело к иному типу структуры, но с вновь открывающимися возможностями

и вызовами. **Неполные формулы** – это **движущая сила семантического анализа** текста.

Цель анализа – методом взаимного заполнения неполных формул собрать целые Ситуации и доказать **связность текста** (диагностируя места и степень несвязности). При выбранном способе единообразных двуместных формул можно дополнять или сокращать СемПространство, не ломая целой структуры. Семантический граф всего текста будет неизбежно **сжатой структурой**, так как в него должны войти только полноценные единицы. А оставшиеся в нём неполные формулы – сигналы о том, что можно или нужно выходить в межтекстовое пространство…

5. Об Информационном пространстве текстов

В каком виде мы можем добавить в СемПространство новый компонент – пользователя, ищущего свой Смысл в тексте? Пользователь может войти в систему лишь в виде заданного им текста запроса или вопроса. Он назван «**встречным текстом**» и обрабатывается стандартными процессорами, вокруг него как ядра должен формироваться ответ Системы.

Текст вопроса вольётся в СемПространство текста, не нарушая свойства непрерывности. Пользователь может усилить или даже заменить вопрос добавлением других «встречных» текстов. К ним относятся: а. дополнительный текст, уточняющий Ситуацию, описанную в вопросе; б. спецзнания, они же задают единицы и терминологию, в которых желательно получить ответ.

Добавленные источники или построенные из них структуры, будучи включёнными в СемПространство, получают заведомо больший информационный ВЕС, чем текстовые единицы, так как окончательный ответ Системы должен быть ориентирован на их лексику и грамматику. Семантическое Пространство, в которое добавлены любые внешние Знания, становится **Информационным пространством** этого текста[11]. Вычисление новой Информации, или **индивидуального Смысла**, на структуре Семантического или Информационного Пространства требует новых методов работы; это вызов программистам, лингвистам и логикам.

Добавление усложняющих модулей (текст пользователя, встречные тексты как Знания и неполные формулы как Незнания), расширяющие исходную структуру до Пространства (Семантического или Информационного), может дать новый импульс развитию системы Смыслов и Текстов, так как нацелено на актуальные практические интересы пользователей.

6. Зачем нужна другая модель

Оставаясь в рамках общей концепции «Смыслов и Текстов», вложим в понятия **Смысл** и **Текст** более «человеческое» понимание. Предлагается более мягкая и более широкая трактовка Текстов и Смыслов, названная нами моделью АПТ (автоматического понимания текстов). Сама заявка на моделирование **понимания** требует выхода за пределы чисто лингвистических структур и процессоров, хотя бы по той одной причине, что только лингвистический анализ НЕ включает субъекта процесса понимания (=самого «понимателя», или пользователя, или читателя текста), хотя это самый активный компонент Системы. Главной функцией системы АПТ будем считать вычисление новой для читателя **Информации** из данного текста. Читатель воспринимает Текст как целое образование со всеми его стандартными параметрами, как то: Авторство текста, его Специализация, самые важные понятия и утверждения и др. Его также интересует, в какой массив (сборник, книгу, газету) входит статья или произвольный фрагмент текста. Все параметры текста (когда статья появилась, её размеры, наличие схем и иллюстраций и т.д.) имеют Смысл для него и должны как-то отобразиться в структуре. Только увидев полную картину текста, человек или отбросит текст как бесполезный для него, или сформулирует (тоже в виде текста) то, что является для него искомой информацией. Ведь главный его интерес состоит в том, чтобы найти такую порцию сведений в тексте, которая согласуется с его информационным запросом. Естественно, что другой читатель построит из того же текста другой СемП, или свой **собственный Смысл**.

Иначе говоря, **Смысл** в модели АПТ – не застывшая структура (СемП=Смысл), а функция, вычисляющая Смысл, который строится динамически, учитывает всех участников процесса понимания с их параметрами. Модель и систему, нацеленных на получение полезной Информации, обычно относят к информационно-поисковым системам (ИПС). Но правильнее считать её **информационно-лингвистической** моделью: роль лингвистики в ней очень важна.

Способ совмещения этих двух дисциплин – Лингвистики и Информатики, который бы оправдал использование термина «автоматическое понимание», пока не найден. Система машинного перевода (МП) как самая представительная лингвистическая система и ИПС как типичная система разных поисковых процессов пока НЕ нашли дороги навстречу друг другу. Между тем обе дисциплины нужны: если ИПС помогает пользователю отобрать корпус потенциально интересных (=релевантных) текстов, то лингвистическая система отвечает за адекватность анализа текста и способность построить краткий правильный (и содержательный) ответ на любой запрос.

В работе автора «Автоматическое понимание текста»[12] рассмотрены системы типа IE (Information Extraction). Хотя в них совмещены Информатика и элементы Лингвистики, в них те же недостатки, что и отдельно в системах МП и ИПС. Они полезны для очень конкретных производственных задач типа «Сколько человек уволилось», «Сколько человек поступило на работу», «Какова прежняя работа начальника» и т.п. Они моделируют «механическое понимание», опирающееся на лексические и структурные повторы с текстом. Такие системы вряд ли смогут включить, скажем, формальный синтаксический компонент или построить на выходе грамматически правильный ответ.

7. Почему удобен термин «понимание»?

Слово и понятие ***Смысл*** как окончательное СемП очень категорично, оно – как выстрел: хочешь – не хочешь, а его надо достроить, причём оно должно быть однозначным, формальным, точным и т.д. А ведь огромное количество текстов вообще не имеют смысла, а представляют собой пустое перебирание слов, демагогии и пр., и это видно с первого взгляда любому человеку.

Термин ***понимание*** намного мягче, он допускает любое частичное понимание: можно понять и отдельное слово, и лишь несколько слов (*через пень колоду*), и отдельную реплику, но всё же уловить, понять, какую информацию хочет передать автор. Понимание может быть буквальным, лобовым, либо изощрённым, когда все слова восприняты в их буквальном виде, но целое из них «не складывается». Оно может быть приблизительным, чего нельзя сказать о классическом СемП. Построением промежуточного СемП в модели АПТ процесс поисков смысла только начинается. Оно заканчивается, если и когда доказана **связность** анализируемого текста.

Лингвистическое СемП – это, по замыслу, **объективная** структура: при едином аппарате и правилах построения должен получаться для заданного объекта один и тот же результат. Понимание же в принципе **субъективно**, индивидуально: при каждом новом запросе (встречном тексте) будут строиться разные ответы. Регулируя и уточняя свой вопрос, пользователь может получать в каждом акте работы с системой всё более точные результаты. Естественно, что для разных пользователей, имеющих разные цели при поиске нужной им информации, и при задании разных «встречных» текстов, система АПТ будет строить разные ответы. Наш **смысл** (в предлагаемой модели) всегда **индивидуален**.

Итак, мы моделируем «мягкое» понимание текста. Это означает, что структура СемП должна быть **адаптивной**. Результат должен

вычисляться каждый раз по-новому, по формуле, в которой мы можем менять количество и состав компонентов. Так, можно включать и выключать компонент внешних знаний или какую-то его часть. Можно убрать один из лингвистических процессоров (например, анализирующий фразеологизмы или делающий первичный разбор текста); при анализе очень плохих массивов можно исключить даже синтаксический анализ предложений и оценить результат – как понята морфология словоформ, какую интерпретацию она получила на первичном семантическом уровне. Если же ставится задача оценить грамотность построения фраз автором, можно оставить синтаксический анализ, отключив семантическую интерпретацию. На каждом шаге работы процессор может быть остановлен и запрошены промежуточные итоги с оценками тех этапов, которые проработали.

Модель **«Смысл⇔Текст»** объявлена способом формального описания **языка**, а модель АПТ по своему названию принадлежит к прикладным системам обработки и «понимания» текстов, или **речи**. Насколько первая выполняет своё теоретическое назначение, мы обсуждать не будем, это задача высокой литературы, но у прикладной системы понимания текста должна быть своя теория, объясняющая, какие процессы человеческого интуитивного восприятия она имитирует.

Прикладная теория не обязана совпадать с лингвистическими теориями. Хотя Язык и Речь (Текст) тесно связаны как двуединая сущность, это две разные системы. Чтобы сделать шаг вперёд относительно способов понимания естественных текстов, учитывая то, что дали лингвистические разборы предложений, необходимо в полной мере опираться на законы построения самих текстов.

8. Проблема специальных Знаний в модели АПТ

Система понимания и получения информации из текста, а также машинный перевод, нужны прежде всего специалистам в самых разных предметных областях. Поэтому из всех построенных системой Текстовых структур[13] и «встречных» текстов самыми важными будут тексты или их фрагменты, близкие к профессиональным интересам читателя. Но как найти нужную ему базу знаний – неужели перебором всех имеющихся «спецбаз» данных и знаний? Конечно, в Систему АПТ желательно привлекать максимально формализованные источники спецзнаний. Но это практически невозможно, потому что и у самих баз данных и/или знаний нет общего языка обмена информацией, хотя таких попыток сделано много. А главное, нет языка

общения произвольной БД с произвольным текстом[14]. Проблему совместимости разных Баз данных и знаний, их сравнения и т.п. вряд ли можно будет решить только программистскими и технологическими приёмами. Научить разные БД экономно разговаривать друг с другом скорее могут лингвисты, предложив язык, адаптируемый к самым разным формам фиксации знаний, то есть своего рода **информационный язык-посредник (ИЯП)**. Процессоры, устанавливающие связи с ВнеТекстовыми источниками Знаний и использующие эти знания в анализе текста, суть системы «машинного перевода» с одного языка на другой, будь то один естественный язык или разноязычные источники. Для лингвистов и систем МП это очередной «иностранный» язык, а задача по сути лингвистическая: в ней те же проблемы, что и в машинном переводе, главная из них – создание информационного языка-посредника[15].

Нужен упрощённый смысловой метаязык, адаптируемый к разным Предметным областям, – своего рода лингвистический Ассемблер, позволяющий сравнивать и дополнять СемПространство текста нужными знаниями. Наши эксперименты с разными прикладными системами и моделями сводились в основном к уточнению «верхнего уровня» такого адаптивного метаязыка. Необходимое его обогащение и специализация должны происходить в процессе адаптации к разным способам фиксации Знаний в разных науках. Кстати, такой же ИЯП позволит **преодолевать стыки** и между единицами лингвистических структур разных уровней, прежде всего синтаксической (или СинП) и семантической, даже если они неполные: ведь в модели АПТ единицы синтаксические и семантические – это единицы разной природы. Но эта тема требует отдельного рассмотрения.

9. Взаимодействие уровней понимания в системе АПТ

Итак, в качестве **объекта** исследования мы выбираем любой естественный текст (ЕТ) в электронном виде, независимо от его происхождения, от качества, величины и т.д. Лингвисты и их системы умеют строить разные его отображения в виде четырёх последовательно получаемых структур: первичной (выделяющей сами единицы в составе текста, в русской традиции его называют Графематическим анализом), морфологической, синтаксической и даже первичной семантической интерпретации (СинСемП). Каждая из них может быть безупречной, но может содержать ошибки или разные виды дефектов.

Дефекты более раннего уровня анализа часто могут быть сняты следующими за ним структурами. Так, если Морфология выделила в потоке текста две единицы (***в*** – как предлог и следом за ним слово ***ведении*** как существительное), то Синтаксис соединит их, назвав всё словосочетание *в ведении* сложным предлогом: А ***в ведении*** В, или ***В-ведении*** (А,В). Каждый уровень может ошибиться, но он может обратиться к следующим уровням анализа или «посоветоваться» с каким-то внешним источником: с разными Словарями, например, словарём фиксированных оборотов, статистиками или прочими привлекаемыми помощниками, часто это стандартная Грамматика естественного языка. Обращаться можно и к Грамматике текста, если она постепенно формируется по ходу анализа одного или некоторого множества текстов. Но Семантика имеет право ещё раз пересмотреть это решение, если оно не согласуется с законами правильной семантической структуры.

Уже из такого краткого упоминания сути лингвистической работы с текстом можно понять, что это сложная и динамическая задача. В ней предусмотрены и Диалоги (когда сравниваются разные структуры), и выводы логического или статистического характера, и возвраты (когда надо исправить ошибки раннего слоя анализа, скорректировав затем и следующие за ним структуры), и даже собственно «машинный перевод». Обмен информацией, или диалог уровней анализа, если таковой предусмотрен в Системе АПТ, моделирует и Диалог следующего уровня, когда в работу подключается следующий Компонент – реальный Пользователь – ставящий свою цель и желающий извлечь из текста свой индивидуальный Смысл.

Кроме того, в любом тексте незримо присутствует (точнее сказать, отсутствует в материальном виде) Нечто, что принято называть законами Действительности. Частично это проявляется в нарушении семантического согласования при анализе, что фиксируется явно в структуре как одна из разновидностей смысловой неполноты. Действительность – это ведь в буквальном смысле «самая Предметная» область знаний из всех научных областей. В предложенной модели АПТ отсутствующее знание о ней может задаваться, пока не найдено лучшее решение, в виде короткого текста подсказки (например, «Птицы в норме имеют два крыла», «Люди обычно имеют два глаза, две руки» и т.п.), если в тексте упомянут соответствующий объект. Подсказка тоже переводится в структуру нашего первичного СемП и потом пополняет Информационное пространство как очередной «встречный» текст, помогая разрешать трудные случаи локальной омонимии.

Считаю, что минимальная сфера действия Семантики – целый текст, а максимальная – Информационное пространство текстов.

Ведь сравнение содержания анализируемого текста с другими текстами на ту же или другую близкую тему (с вычислением степени похожести, вплоть до заимствования) никак нельзя исключать из «Семантики» в самом высоком понимании этого термина. Именно такая Семантика объединит «Прикладную лингвистику» с системами типа ИПС, поэтому модель АПТ мы и назвали **информационно-лингвистической** моделью.

Выводы:

1. Три главных трудности в общей проблеме Автоматического понимания текстов (АПТ) – неоднозначность, неполнота и отсутствие семантической связности. Именно эти три явления неопределенности привели в свое время автора статьи к нестандартному представлению сначала лингвистических структур предложения[16], а впоследствии и структуры целого текста[17].

2. Предложенная «мягкая» модель АПТ объединяет теории естественного понимания текстов с «машинным переводом». Главное, что она моделирует **множественное понимание** одного текста разными воспринимающими устройствами, в том числе реальными пользователями. Мы рассматривали в основном трудные и «плохие» тексты, коими полны СМИ: в них явления неоднозначности и неполноты особенно очевидны. Движущей силой в структурах текста (его Семантического Пространства) является Вопрос, неполнота. Ликвидация неполноты приводит к расширению, а затем и **сжатию** структуры за счёт устранения повторов ситуаций и их частей. Образование более сложных и более содержательных единиц тоже сжимает структуру. При этом всегда происходит **скачок**, изменение Языка общения. Сжать текст так, чтобы изложить его содержание короче, и означает **понять его содержание**. Желательно, чтобы лингвистический Синтез мог строить правильные тексты из структур с разными степенями сжатия, включая разные табличные формы и БД. Представляется, что для систем автоматического синтеза гораздо легче реализовать перевод сжатых форм представления содержания текста (если, конечно, они оценены как адекватные тексту и заслуживают того, чтобы их вообще переводить).

3. Работы по созданию систем машинного перевода развернулись в 60-х годах прошлого века; их первые результаты были оценены специальной комиссией в США (ALPAC). Вывод, сделанный этой комиссией, был строг: промышленный и высококачественный МП недостижим[18]. В настоящее время «промышленные» системы МП и

ИПС пока реализуют перевод, близкий к пословному, и квазиреферирование-аннотирование соответственно, но и такие результаты важны. Только коллективы с серьёзной поддержкой да группы энтузиастов машинного перевода остались верны лингвистическим теориям, – очень важно, что они также формируют и уточняют саму теорию по ходу продвижения и реализации работ[19].

4. У прикладных систем обработки текстов должна быть **своя модель и теория**, точнее, своя «философия» – она и требует переопределения классических понятий модели «**Смысл⇔Текст**». В модели АПТ входные тексты могут быть произвольными: любой отрезок текста, даже если он без начала и конца, может быть проанализирован и оценен как нормальный или дефектный на разных уровнях «понимания». Результат понимания в модели АПТ представляет собой сложный мультиграф: он моделирует множественное и неоднозначное понимание текста разными пользователями и в разные стадии «роста» как текстовых структур, так и самих пользователей.

5. «Дефекты» текстовых сообщений (неполнота в широком смысле), отражаемые в явном виде в структурах, используются в конструктивном ключе: движущей силой семантического анализа является сам Вопрос и неполные участки построенных локальных структур. Взаимное заполнение неполных формул в семантическом и информационном пространстве текста позволит не только восполнить недостающие на локальных участках знания (в том числе внетекстовые), но и приводит к сжатию структуры. Сжатие (или содержательная компрессия) текста происходит и за счёт вытеснения маловажных участков содержания, а этой операции предшествует оценка информационного ВЕСа единиц СемП как промежуточной структуры. Построение более содержательных единиц – всегда Скачок навстречу единицам заданных «встречных» текстов. Экспликация неоднозначности расширяет СемПространство и моделирует **множественность** субъективных «пониманий» текста построением **индивидуальных смыслов**.

6. Хотя реализовать полностью автоматическую модель АПТ не представляется реальным, необходимо использовать в ней максимально продвинутый лингвистический анализ. Однако самые серьёзные проблемы естественного понимания лежат вне собственно лингвистической теории. В АПТ они начинаются с этапа «первичного разбора» текста и продолжаются в режиме «далее везде». Особенно «богаты неполнотой» также последние структуры, включающие знания адресата информации и другие «встречные» тексты.

7. Ответ в виде структуры индивидуального знания использует пока только человек, но в полном виде системы могут быть предло-

жены разные пути продолжений, например: передать полученную структуру следующему этапу понимания, включить её в компонент Общих Знаний, оценить текст как «плохой» или даже «вредный» для общества и т.д. Модель АПТ, частично реализованная в системе французско-русского автоматического перевода ФРАП, с самого начала включала компоненты Автора и Адресата, а также диалог как со всеми уже построенными структурами, так и со специальными словарями, Тезаурусами и базами спецзнаний[20].

8. Мы видим в таких «мягких» структурах АПТ возможность моделировать и некоторые социальные процессы, например, процессы образования классов и групп и их взаимодействий в информационном обществе, «обратную связь» между обществом и некоторыми подструктурами. В нашем социуме конфликты и непонимания достигли высшего уровня, и не исключено, что работающее устройство АПТ сможет подсказывать правила выхода из конфликтных и тупиковых ситуаций[21]. При этом на структурах Семантического и Информационного пространств текста дистрибутивно-статистический метод А.Я. Шайкевича сможет работать эффективнее, чем на «сыром» тексте, если он будет считать повторы не только лексем, но и Ситуаций и их частей. (Рассмотрение и обоснование других следствий из модели АПТ требует экспериментов).

9. В системах АПТ метаязык понимания текста (ИЯП) не надо создавать «с нуля»: во многих ИПС, а также в базах данных и знаний, анкетах, таблицах элементы такой искомой грамматики так или иначе присутствуют. Что касается их состава, то они имеются в большом количестве в любых НИИ, организациях и ведомствах, технических и гуманитарных (Номенклатуры, Словари, Тезаурусы, действующие информационно-поисковые системы, Отчёты и просто корпусы разных протоколов и текстов). Но их важно выявить путём лингвистического анализа тех текстовых элементов в Базах данных и знаний, которые составляют собственно содержание знания.

Эксперименты с полу-ручной обработкой текстов, включая работающие автоматические компоненты, вполне реально проводить силами даже студентов, если в них (экспериментах) поставить задачу сбора кандидатов смысловых единиц, предлагаемых в Список **базовых** смысловых отношений как основу для семантического метаязыка. Следующая не менее творческая задача – **адаптация**. Она состоит в создании правил укрупнения базовых (всегда двуместных) отношений Семантического пространства текста, или нашего промежуточного СемП, в сложные предикаты, разные для разных специальных Знаний.

10. Работа в Пространстве текстовых структур даст много полезного гуманитарным наукам, напр., «Лингвистике текста»[22], психолингвистике и другим когнитивным дисциплинам.

Примечания

1 *От редколлегии.* Статья Н.Н. Леонтьевой, обобщающая многолетний опыт работы автора, посвящена трудностям и особенностям применения заглавных понятий теории моделей «Смысл⇔Текст» в прикладных разработках. В этом же номере нашего журната помещена 2-я часть фундаментального исследования И.А. Мельчука, публикацией которого редколлегия отметила 50-летие теории моделей «Смысл⇔Текст». Думается, что страстный полемический накал статьи Н.Н. Леонтьевой – лучшее свидетельство жизненности и актуальности затронутых обоими авторами научных проблем. Редколлегия сочла важным сохранить особенности авторского языка и стиля изложения, в том числе написание вводных фундаментальных понятий с заглавных букв.

2 Основой данного текста был доклад автора «Семантические тупики в системах автоматического понимания текстов» на конференции в Санкт-Петербурге, он опубликован в сборнике избранных научных статей «Инфраструктура научных информационных ресурсов и систем» (*Леонтьева Н.Н.* Семантические тупики в системах автоматического понимания текста // Пятый Всероссийский симпозиум «Инфраструктура научных информационных ресурсов и систем» (5–8 октября 2015 г.) / ВЦ РАН. С.-Петербург, 2015. С. 109-121.) Если доклад был обращён к специалистам по информационным системам и способам представления Знаний, то настоящее изложение адресовано молодым лингвистам, от которых я жду дальнейших исследований и уточнений структур Информационного Пространства текстов. Именно на этом пункте наша работа над системой ФРАП была остановлена в связи с наступлением эры персональных компьютеров. Дальнейшая работа с «трудным» текстовым материалом шла «устно», на семинарах со студентами кафедры ТиПЛ, которые с удовольствием расписывали разные структуры, включающие неполноту, и заполняющие её «встречные» знания. Всем участникам экспериментов моя благодарность. Я благодарю также Сергея Иосифовича Гиндина за помощь в редактировании данного текста и Н.Г. Семёнову за помощь в оформлении Библиографии, см. Материалы к библиографическому указателю печатных работ Н.Н. Леонтьевой / Сост. С.И. Гиндин, Н.Г. Семенова // Вестник РГГУ. Серия «Языкознание». 2007. № 8. С. 215-235 (Московский лингвистический журнал, Том 9/2).

3 *Mel'cuk I.* A General Inventory of Surface-Syntactic Relations. Part One // Вестник РГГУ. Серия «История. Филология. Культурология. Востоковедение». 2015. № 8. С. 75-103 (Московский лингвистический журнал, Том 17 (2)). Считаю очень ценным публикацию полного перечня поверхностно-синтаксических связей, так как они «более семантичны», чем так называемые глубинно-синтаксические, об этом мы говорим в книге «Автоматическое понимание текста» (см. примеч. 4 ниже).

4 *Леонтьева Н.Н.* Автоматическое понимание текста: Системы, модели, ресурсы : Учеб. пособие. М.: Академия, 2006. 300 с.

[5] См. *Мельчук И.А.* Русский язык в модели «Смысл⇔Текст», М. Вена, 1995; *Апресян Ю.Д.* Интегральное описание языка и системная лексикография // Избранные труды. Том 2. М. : Языки русской культуры, 1995. Это лишь две произвольно выбранные публикации.

[6] См. работы разных лет по синтаксису предложения Л.Н. Иорданской, а также книгу: *Кобзарева Т.Ю.* В поисках синтаксической структуры. Автоматический анализ русского предложения с опорой на сегментацию. М., 2015. 377 с. Иорданская и Кобзарева вслед за Л. Теньером называют Ситуацией собранную в пределах предложения главную подструктуру (по заданной «модели управления»). Такая синтаксическая ситуация как правило семантически неполна.

[7] Слишком много работ и дискуссий, особенно в ранние годы МП, было посвящено теме Языка-посредника для машинного перевода, в частности, они отражены в выпусках сборника «Машинный перевод и прикладная лингвистика» (МП и ПЛ), а также в Препринтах группы ПГЭПЛ ИРЯ РАН.

[8] См. сборник Машинный перевод и прикладная лингвистика. Проблемы создания системы автоматического перевода: Сб. науч. трудов МГПИИЯ им. М. Тореза. М., 1987. Вып. 271, в котором описаны разные аспекты работы системы ФРАП, а также статью в нём: *Леонтьева Н.Н.* Система французско-русского автоматического перевода (ФРАП): лингвистические решения, состав, реализация // Там же. С. 6-25.

[9] *Леонтьева Н.Н.* О смысловой неполноте текста (в связи с семантическим анализом) // Машинный перевод и прикладная лингвистика, вып.12. М., 1969. С. 96-114.

[10] *Сокирко А.В.* Реализация первичного семантического анализа в системе Диалинг // Труды Международного семинара «Диалог 2000» по компьютерной лингвистике и её приложениям. (Протвино, 1-5 июня 2000 г.). М.: Наука, 2000.

[11] Более подробно см. статью автора: *Леонтьева Н.Н.* Об информационном пространстве текстов // Znaczenie. Tekst. Kultura. Tom 5. Warszawa 2014. С. 371-384.

[12] *Она же.* Автоматическое понимание текста ...

[13] *Она же.* Построение Базы текстовых фактов // НТИ. Сер. 2. М., 1990. № 7.

[14] *Рубашкин В.Ш.* Онтологическая Семантика. Знания. Онтологии. Онтологически ориентированные методы информационного анализа текстов. М., Физматлит, 2012. 348 с. См. также указанную в книге литературу по языкам представления Знаний (ЯПЗ).

[15] Укажу лишь некоторые публикации автора: *Леонтьева Н.Н.* Создание информационного языка на базе семантического анализа текста // НТИ. Сер.2. М.,1971, № 8; *Она же.* Семантика связного текста и единицы информационного анализа // НТИ. Сер. 2. М., 1981. № 1; и другие.

[16] *Она же.* Об одном способе представления смысла текста // Информационно-поисковые системы и автоматическая обработка научно-технической информации. Том 2, 1967. С. 192-204.

[17] *Она же.* Корпусная лингвистика: не только вширь, но и вглубь // Труды Междунар. конференции «Корпусная лингвистика-2006». СПб: Изд. СПб ун-та, 2006. С. 234-241 (в работе предложены способы представления и именований сжатых текстовых структур Ситуация-Событие-Текстовый факт); *Она же.* Построение Базы текстовых фактов // НТИ. Сер.2. М., 1990. № 7.

[18] В России плановые работы по МП были заморожены на 10 лет, а коллективы лабораторий МП в Москве, Ленинграде, Киеве, Ереване и многих других городах были переориентированы на создание других систем – ИПС, АСУ и подобных им (в ЛМП МГПИИЯ шутили: «Будем теперь создавать АСУ кашля и насморка»). Переходить на информационные системы было вполне закономерно, ибо и в системах МП был необходим поиск информации. Работа над разными ИПС не была бесполезной: она помогла расширить горизонты, уточнить будущее «идеального» машинного перевода. А МП как наука до сих пор представляет собой прекрасный и богатый образец исследовательской деятельности в области создания так наз. «искусственного интеллекта». Хотя комиссия ALPAC работала 50 лет назад, её выводы верны и сейчас: идеальный машинный перевод в обозримом будущем недостижим как для реальных специальных текстов, так и для художественных, доводить текст до кондиции должен всё же человек. Исключения крайне редки: для очень ограниченных структурно и лексически корпусов единообразных текстов, таких как прогнозы погоды в переводах системой METEO, работающей в Канаде на основе МСТ.

[19] В России это в основном коллективы, работающие под руководством академика Ю.Д. Апресяна, как наиболее последовательные продолжатели модели и лингвистической теории, заложенной в классических работах Мельчука и Апресяна (примеры см. выше).

[20] Без них никакой практически значимый машинный перевод пока невозможен, так как он в основном и востребован специалистами в разных областях науки, далёких от лингвистических проблем, особенно в такой организации, как ВЦП, которая ежедневно получает «чемоданами» заказы на перевод статей по самым разным, порой экзотическим тематикам, как разведение пчёл в древнем Израиле.

[21] Более развёрнутые описания и обоснования структур АПТ можно найти на сайте МГУ http://leontyeva.srcc.msu.ru/, который медленно заполняется (метод сканирования и коррекции старых, докомпьютерных работ очень трудоёмок), а в самих статьях есть ссылки на аналогичные подходы к АПТ и примеры, чему нет места в данной, по сути обзорной, заметке.

[22] См. *Гиндин С.И.* Онтологическое единство текста и виды внутритекстовой организации // Машинный перевод и прикладная лингвистика. 1971. Вып. 14. С. 114-122; *Гиндин С.И., Леонтьева Н.Н.* О понятии “текст” // Проблемы анализа и синтеза целого текста в системах машинного перевода, диалоговых и информационных системах / Всесоюзный Центр Переводов. М., 1978. С. 75-83. (Сер. 2 : Машинный перевод и автоматизация информационных процессов. Обзорная информация); *Леонтьева Н.Н.* Три свойства связного текста // Tekst i zdanie / Pod red. T.Dobrzinskiej i E.Janus. Wroclaw e.a.: Ossolineum, 1983. S. 171-181; Ср. также работы польской научной школы по указанной теме: *Янус Э.* Обзор польских работ по структуре текста // Синтаксис текста / Отв. ред. Г.А. Золотова; ИРЯ АН СССР. М.: Наука, 1979. С. 325-340.