ХАРАКТЕР КОРРЕЛЯЦИИ МЕЖДУ ПОРЯДКОМ СЛОВ И КОММУНИКАТИВНОЙ ПЕРСПЕКТИВОЙ В НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТАХ НА РУССКОМ ЯЗЫКЕ

CHARACTERISTIC RELATIONS BETWEEN WORD-ORDER AND COMMUNICATION PERSPECTIVE PATTERNS IN RUSSIAN SCIENTIFIC TEXTS

Гордеев С. С*. (propogss@mail.ru),* ***Азарова И. В.*** *(azic@bsr.spb.ru)*
Санкт-Петербургский государственный университет

В статье рассматриваются типы соотношений моделей расположения субъекта, предиката, объекта в простых предложениях и актуального членения предложений, которые будут использоваться при определении зон вероятного размещения новой и актуализированной информации в грамматической структуре предложений в формально-грамматическом парсере Russ4IR.

Параметры порядка слов исследовались по выборочной совокупности из корпуса современных текстов кафедры математической лингвистики СПбГУ. Устойчивые параметры коммуникативной структуры предложений были выявлены при ручной разметке совокупности научно-технических текстов. Были определены ядерные и периферийные компоненты коммуникативной структуры, а также доминирующие схемы соотношения словопорядка и тема-рематического членения предложений текста.

1. Введение

При автоматическом грамматическом разборе текста в парсере Russ4IR[1], использующем формальную грамматику AGFL[2], определяются различные параметры структуры простых предложений: базовый тип структуры (глагольный или именной), структура словосочетаний, занимающих позиции главных членов предложения и их распространителей, порядок слов (объективный, субъективный или смешанный). Полученная грамматическая структура в дальнейшем может интерпретироваться семантически. Однако для некоторых задач полный семантический анализ, возможно, будет слишком «затратной» процедурой. Исходя из предположения, что содержание текста может передаваться как в виде набора пропозиций, так и в виде «размеченных» синтаксически связанных групп слов, представленных в нормализованном виде[3], мы предлагаем в данной статье варианты характеристик грамматической структуры, которые связывают линейную (линейно-синтаксическую) структуру текста с актуальным членением предложений текста. Хорошо понимая, что предлагаемые нами описания носят в большой степени предварительный характер, мы будем рассматривать эти варианты применительно к научно-техническим русским текстам, которые в значительной степени лишены экспрессивной нагрузки.

Актуальное членение (АЧ) суть явление, являющееся фактом мышления и психологии. Очевидно, что АЧ существует во всех естественных языковых системах и осуществляется посредством набора языковых средств, но столь же очевиден и его «надсистемный», субъективно-ментальный статус. Особенно следует подчеркнуть текстообразующую функцию коммуникативной организации высказывания, которая осуществляет связь эксплицитного текста, доступного интерпретации адресата, с «имплицитными сообщениями: пресуппозициями и постсуппозициями»[4]. Однако на текстовом уровне можно говорить о субъективном характере глобальной коммуникативной организации текста, поскольку последовательность тематических структур текста выделяется

[1] Азарова И.В., Иванов В.Л., Овчинникова Е.А. Использование схемы наследования рамок валентностей в тезаурусе RussNet для автоматического анализа текста // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006». М., 2006. С. 18-25.

[2] URL: http://www.cs.kun.nl/agfl/

[3] Phrase-based High Accuracy Search, Analysis and extRaction of Metabolite data from Literature // URL: http://www.phasar.cs.ru.nl/

[4] Ступина И.Ю. Коммуникативная организация высказывания в контекстном и семантическом аспектах: Автореф. дис. … канд. филол. наук. Л., 1989.

субъективно[5], т. е. текст как таковой не обладает четко выраженной макроструктурой, а она приписывается ему читателем или автором текста.

Считается, что основным средством выражения АЧ в современном русском языке является интонация и порядок слов[6]. Помимо этих средств, могут использоваться также частицы и модальные слова. В письменных научно-технических текстах порядок слов приобретает особое значение в силу «информативного» характера текста, зачастую не предполагающего устной реализации.

2. Основные схемы словопорядка в современных русских письменных текстах

Русский язык обычно характеризуется как язык со относительно свободным порядком слов (freer word-order language). Насколько справедлива такая характеристика применительно к письменным текстам, в которых только предположительно могут использоваться интонационные средства?

Чтобы проверить, имеются ли доминирующие схемы словопорядка в структуре простых предложений, мы исследовали выборочную совокупность в 1000 предложений из корпуса современных текстов кафедры математической лингвистики СПбГУ «Бокренок»[7]. В эту совокупность были включены случайным образом предложения, представляющие разные функциональные стили. Поэтому предполагалось, что схемы словопорядка в исследуемой совокупности будут давать более разнообразные структуры, чем в той области, которая является объектом изучения в данной статье.

Поскольку в синтаксическом модуле Russ4IR мы используем идею «топологических полей»[8], когда задается в качестве доминирующего признака положение субъекта и объекта относительно предиката, мы рассмотрели частотность реализации схем объективного (прямого) и субъективного (обратного) порядка слов, которые представлены на схеме 1. Оказалось, что объективный порядок наблюдается в 73.41% предложений, субъективный – 20,3%, смешанный – 5,89%, одиночные предикаты – 6,64%.

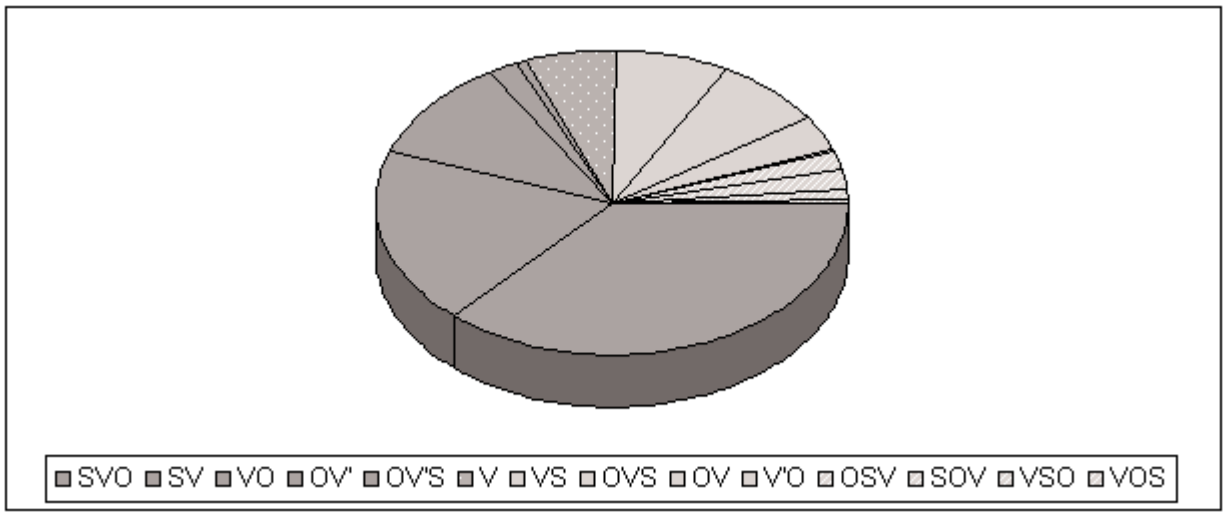

Схема 1. Распределение схем объективного, субъективного и смешанного словопорядка в простых предложениях в выборочной совокупности корпуса

Таким образом, корпусные данные показывают, что доминирует объективный порядок слов, наиболее частотной является схема словопорядка SVO, затем SV и VO. Значимую частотность (>5%) имеют схемы субъективного порядка VS и OVS и V. V' используется для пассивных форм предикатов, причем общее количество пассивных форм составляет лишь 3% от общего числа предложений. В выборочной совокупности имелось столь незначительное количество (<3%) вопросительных предложений, что они не представлены в сводном описании.

---

[5] ван Дейк Т.А. Язык. Познание. Коммуникация. М., 1989.

[6] Русская грамматика. Т. 2. М., 1980. С. 191.

[7] Азарова И.В., Синопальникова А.А. Использование статистико-комбинаторных свойств корпуса современных текстов для формирования структуры компьютерного тезауруса RussNet // Труды международной конференции «Корпусная лингвистика 2004». СПб., 2004. С. 5-15.

[8] Penn G., Haji-Abdolhosseini M. Topological Parsing // Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03), Budapest, Hungary, April 2003.

Соотношение частот встречаемости схем словопорядка объективного и субъективного типа напоминает сходные соотношения, которые мы наблюдали при обследовании распределения контекстов значений полисемантичных слов в корпусе «Бокренок»[9]. Отсюда частотность схем можно интерпретировать следующим образом: объективный словопорядок имеет стереотипный характер и проявляется столь же регулярно, как основные значения слов; субъективный словопорядок имеет контекстно-зависимый характер и может быть ассоциирован с некоторыми сложными конструкциями; пассивные конструкции и смешанный словопорядок являются окказиональными.

2.1. Схемы словопорядка в русских научно-технических текстах

При разметке научно-технических текстов мы проверили распределение частот выделенных схем словопорядка и получили практически такие же результаты по соотношению объективного/ субъективного/ смешанного порядка: 76,6% - 18,4% - 2,5%. В общем, приведенное выше соотношение схем словопорядка реализуется в более чистом виде. Наблюдается еще большее увеличение частотности наиболее регулярной схемы SVO до 48% и повышение доли пассивных предложений до 6%.

3. Устойчивые модели актуального членения в русских научно-технических текстах

3.1. Единицы актуального членения

АЧ использует отношения, выраженные грамматическим членением, в целях коммуникации. В АЧ предложения реализуется отношение субъекта к предмету сообщения как к «известному» или «неизвестному», или как к «данному» и «новому», которые принято именовать терминами «тема» (T) и «рема» (R), либо, в англоязычной традиции, «топик» и «фокус». Существуют разные подходы к определению темы и ремы. Во избежание неоднозначности при нахождении темы и ремы предложения, мы положили в основу идентификации темы понятие определенности: элемент предложения может считаться темой, если он формально определен (назван) в предыдущем контексте. Таким образом, тематическими могут быть анафорические (в широком смысле) компоненты.

С содержательной точки зрения в тексте выделяется общая тема текста, которая последовательно дробится на микротемы, которые в свою очередь могут дробиться далее на более мелкие подтемы, даже внутри отдельных предложений. В состав подразделений темы, как правило, входят уточняющие компоненты, с помощью которых осуществляется дробление основной темы текста. Например, если темой текста являются белки, то подтемами отдельных предложений могут быть *«сократительные белки»* или *«белки, входящие в состав мембран в комплексе с липидами»*. Во избежание путаницы будет в дальнейшем называть тему АЧ топиком, противопоставляя этот компонент описания структуры предложения содержательной теме текста.

Обычное дихотомическое описание АЧ в терминах T-R нам кажется недостаточным для отображения дополнительных «внерематических» компонентов. Такие идеи высказывались прежде И.Ф. Вардулем[10], который предлагал для описания АЧ особый компонент ситуатив и иерархическую систему вложенных структур. При проведении разметки научно-технических текстов мы столкнулись с необходимостью ввести дополнительные «внерематические» компоненты АЧ:

- Ситуативы[11] (S) – определительные или обстоятельственные обороты, которые относятся к ситуации в предложении в целом и задают своеобразное условие, в котором реализуется противопоставление топика-фокуса: *В нашей Галактике в окрестности Солнца масса темной материи примерно равна массе обычного вещества*;
- Модальные рамки и вводные конструкции (M), которые передают предикаты мнения и знания: *Считается, что определенная формула листорасположения характерна для каждого вида растений; Скорее всего, она состоит из новых, не открытых еще в земных условиях частиц;*
- Пояснения к теме или (чаще) реме (E), уточняющий компонент или перечень значений обычно дается в скобках: *Статистическая обработка данных (тест сопряженности хи-квадрат) проводилась в стати-*

[9] Азарова И.В., Иванов В.Л., Овчинникова Е.А. Использование схемы наследования рамок валентностей в тезаурусе RussNet для автоматического анализа текста...

[10] Вардуль И.Ф. О языковых типах в параметре порядка слов // Очерки типологии порядка слов. М., 1989.

[11] Данный элемент АЧ достаточно часто совпадает с таким распространителем структуры предложения в целом, как детермин-нант (см.: Русская грамматика. Т. 2. С. 149-163), однако детерминанты могут выступать в качестве топика АЧ, а ситуативы могут вы-ражаться причастными и деепричастными оборотами, не являющимися детерминантами.

стической среде; *Принято считать, что эти формулы составляют ряд Фибоначчи (1/2, 1/3, 2/5, 3/8, 5/13, 8/21 и т.д.)*;

• Причинно-следственные и другие логические связки (L), осуществляющие внутритекстовую связанность. Некоторые из них также могут выступать маркерами мены топика либо вводить рематические и внерематические компоненты: *Поэтому мы решили изучить формулы листорасположения у разных видов деревьев, кустарников и кустарничков с очередными листьями на массовом материале; Тем не менее, перспектива представляется весьма оптимистической;*

• Фактивы (K) – конструкции вида «предикат» или «субъект+предикат», которые не являются темой или ремой предложения, а представляют некоторое лицо и его отношение (в широком смысле) к предмету сообщения: *Таким образом, мы выясняли формулу листорасположения, которую можно записать в виде дроби, где числитель – число оборотов, а знаменатель – число листьев; Мы вычисляли на побегах этого года формулу листорасположения; Полинг и Кюри установили 2 основных варианта структуры белковой цепи, которые называются α-спираль и β-форма.*

3.2. Регулярность появления компонентов актуального членения

Были размечены вручную фрагменты научно-технических текстов (объемом 500-600 словоупотреблений) из разных частей (начало, середина, конец) текстов в терминах тех компонентов АЧ, которые были предложены выше. На схеме 2 приведено процентное соотношение частотности появления компонентов АЧ в разметке предложений.

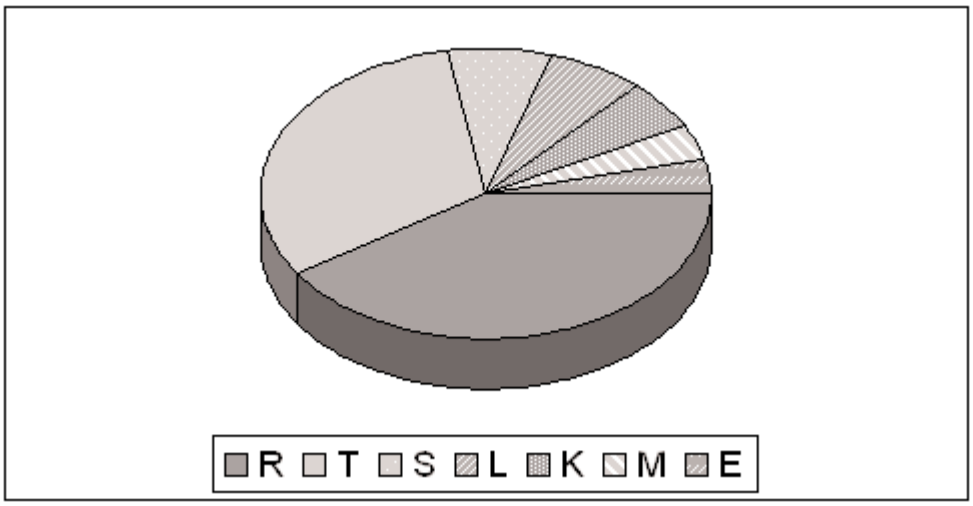

Схема 2. Распределение частотности компонентов АЧ при разметке научно-технических текстов

Распределение частот компонентов АЧ позволяет выделить 3 примерно равные доли: рема (R), топик (T), внерематические (S, L, K, M, E) компоненты. Частота использования компонентов АЧ показывает, что традиционные компоненты T и R являются базовыми, внерематические компоненты – периферийными.

Перечисленные компоненты АЧ различаются не только по частотности использования в размеченных текстах, но также по однократности или многократности вхождения в структуру предложения и их позиционной закрепленности или незакрепленности. Так, базовый компонент R достаточно часто создает структуру с несколькими фокусами:

(1*) Каждая частичка вселенной имеет свое начало и конец, как во времени, так и в пространстве, но вся Вселенная бесконечна и вечна, так как она является вечно самодвижущейся материей.*

Достаточно часто в структуру высказывания входит несколько фактитивов:

(2) *Для этого мы искали листья, расположенные точно друг над другом, считали число листьев между этими двумя листьями, включая один из них, и подсчитывали, сколько оборотов сделает воображаемая спираль, проходящая через основания всех подсчитанных листьев по порядку.*

Напротив, другие компоненты АЧ чаще встречаются однократно и лишь в порядке исключения могут встречаться многократно:

(3) *Белки служат для запасания (примером является миоглобин) и переноса (гемоглобин) кислорода.*

Широко известным фактом является, что компоненты АЧ топик и фокус регулярно занимают соответственно начальную и конечную позиции в высказывании, однако их позиционная фиксированность ниже, чем у такого компонента АЧ, как модальный компонент, которые в 89% своих вхождений стоит в начале высказывания (ситуативы – в 59%, логические компоненты – в 66%).

3.3. Регулярные схемы актуального членения в научно-технических текстах

Полученные при разметке схемы АЧ довольно разнообразны, однако их можно объединить в 4 группы:

- TR предложения и их модификации (70%);
- R предложения и их модификации (16%);
- RT предложения и их модификации (8%)
- смешанные RTR предложения с модификациями (6%).

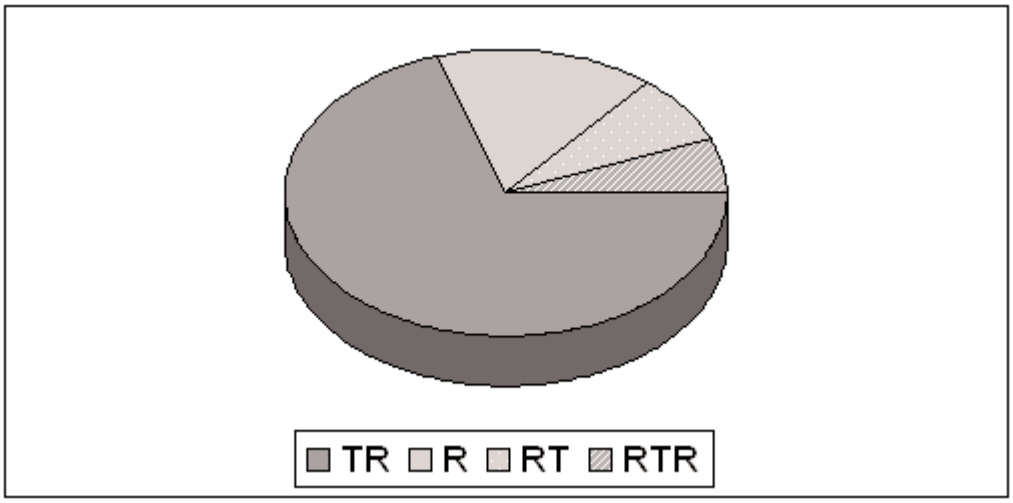

Схема 3. Регулярные схемы актуального членения в научно-технических текстах

Схема TR является самой частотной среди схем АЧ, составляя более трети (35,7%), в первой, наиболее частотной группе она покрывает более половины вхождений. Другие, относительно частотные схемы встречаются на порядок реже: STR – 6%, R – 6%; ITR – 4%; RT – 4%.

Таким образом, в качестве стандартной схемы АЧ следует признать «классическое» TR-членение предложения, которое может модифицироваться внерематическими компонентами. Предложение может содержать несколько TR-групп, которые, как правило, передаются в грамматической структуре в виде простых предложений в составе сложного.

4. Регулярные соотношения между схемами коммуникативного членения и порядком слов в русских научно-технических текстах

Примерно равные доли простых или модифицированных TR-структур в текстах и предложений с прямым порядком слов можно интерпретировать как свидетельство параллелизма этих структур. Действительно, большая часть предложений с объективным порядком слов реализует стандартную TR-схему АЧ, а предложения с субъективным словопорядком – R или RT. Этот достаточно очевидный факт упоминался в большом числе работ, начиная с первых исследований АЧ[12], равно как и регулярная причина использования субъективного порядка – нача-

[12] См., напр., Матезиус В. О так называемом актуальном членении // Пражский лингвистический кружок. М.,1967; Чейф У. Данное, контрастивность, определенность, подлежащее, топики и точка зрения // НЗЛ. Вып.11. М., 1982.

ло обсуждения какой-либо темы, как правило, в начале текста (или абзаца), когда нет реальной темы, еще ничего не было эксплицитно сказано.

С нашей точки зрения, ключевое значение при автоматической интерпретации АЧ грамматической структуры предложения будет иметь определение внерематических компонентов.

Опираясь на проведенные исследования размеченных текстов, мы обратили внимание, что модальные компоненты и фактитивы регулярно используются в начале обсуждения содержательной темы в научно-технических текстах, поскольку необходимо актуализировать «известное», которое, весьма вероятно, входит в систему знаний специалиста. Эта актуализация имплицитной информации воспроизводится лишь в той мере, которая, как считает автор, достаточна для общего указания.[13] Выявление в текстах модальных компонентов и фактитивов в общем не является сложной проблемой, поскольку первые компоненты представляют собой изолированные предикаты или вводные слова, которые можно задать списком, а вторые представляют собой предикаты научно-исследовательской деятельности с субъектами – личными местоимениями или именами собственными, причем сами предикаты поддаются перечислению. Еще один вид фактитива в русских научных текстах – пассивные формы этих предикатов без указания субъекта. После этих компонентов АЧ регулярно используется прямой словопорядок, хотя и вводится новая микротема.

(4) *Так было измерено около 1500 побегов. Далее были построены графики разницы частоты для групп глаголов в позициях окна анализа по которым были определены наиболее характерные позиции для групп глаголов.*

Другая конструкция введения микротемы приводит к изменению словопорядка – это использование ситуативов. После ситуатива регулярно используется обратный порядок слов, причем реализуется R и RT-схема АЧ.

(5) *При объединении аминокислот в белковую цепь образуются пептидные связи (-NH-CO-), на одном конце которых находится NH+3 группа, на другом COO-группа.*

Вариантом предыдущей конструкции будет немаркированная структура, редко встречающаяся в научно-технических текстах – рематическая инверсия.

(6) *Изменялась температура вещества, туманности и состояние, в котором находилось вещество.*

Логические компоненты регулярно встречаются в блоке текста, раскрывающего подтему, т.е. внутри абзаца или даже предложения, в последнем случае вводятся R или TR-схемы АЧ с прямым порядком слов.

(7) *Чем сильнее гравитационное поле, тем быстрее вращаются вокруг галактики звезды и облака газа, так что измерения скоростей вращения в зависимости от расстояния до центра галактики позволяют восстановить распределение массы в ней*

Пояснения в скобках регулярно маркируют рематический компонент.

(8) *Масса нашей галактики оценивается сейчас разными способами и равна $2*10^{11}$ масс Солнца (масса Солнца равна $2*10^{30}$ кг.)*

Особо интересным является деление фрагментов текста, содержащих TR-членение, на формулировку темы (микротемы или подтемы) обсуждения и эксплицитную новую информацию, составляющую научную (в широком смысле) новизну научно-технического текста. Проблема выделения тем, как было выше сказано, зачастую рассматривается как субъективная: автор текста и адресат будут выделять разные фрагменты новой информации. В нашей концепции предлагается использовать формальные признаки фрагментов тем. В первую очередь в качестве тематических маркеров мы предлагаем использовать эллипсис тематических (ключевых, частотных для данного текста) словосочетаний, дополнительно можно использовать местоименные маркеры анафорической замены подчиненных фрагментов словосочетаний. Достаточно часто тематические элементы маркированы лексическими (или корневыми) повторами.

[13] Иногда как известное может приводиться и весьма спорная точка зрения.

5. Выводы и перспективы исследования

Предложенные в статье компоненты АЧ можно использовать для идентификации в документе зон актуализированной эксплицитной информации, то есть формулировок тем обсуждения, которые создают тематический профиль документа, для поиска релевантной информации.

В перспективе в синтаксический модуль предполагается вставить специальный блок для анализа перечисленных компонентов АЧ, что позволит автоматически выделять тематические рубрики в документе в терминах словесных формулировок. Кроме того, с каждым микротемой можно ассоциировать зоны новой информации, не имеющие последующего тематического развертывания, которые будут выступать в роли «новой» эксплицитной информации для профессионального поиска.