

ПРОБЛЕМА КОРЕФЕРЕНЦИИ В РАМКАХ ПОВЕРХНОСТНО-СИНТАКСИЧЕСКОГО АНАЛИЗА РУССКОГО ТЕКСТА

Т. Ю. Кобзарева.
Российский Государственный Гуманитарный Университет.
stamstam@mtu-net.ru

Ключевые слова: автоматический анализ естественного текста, анализ кореферентности, поверхностно-синтаксический анализ.

В работе обсуждаются проблемы анализа отношения кореференции как экстенционального тождества имен в пределах предложения в рамках задач десемантилизованного поверхностно-синтаксического анализа.

Задача - определить характеристические особенности манифестации отношения кореференции в поверхностно-синтаксической структуре русского предложения.

Базисом служит граф межсегментных связей на множестве подграфов синтагматических и сочинительных связей сегментов. Рассматриваются случаи референциального тождества, имплицитруемого наличием синтагматических отношений непосредственно между кореферентными группами или же - выводимого из специфики их места и отношений в структуре графа русского предложения.

1. Введение.

Под кореференцией традиционно понимается референциальное тождество имен [4]. Решение задачи построения отношения кореференции весьма значимо при решении различных задач, связанных с автоматической обработкой естественного текста, так как обеспечит возможность поиска имеющейся в анализируемом тексте информации об экстенционалах имен.

Кореферентность интенционалов предполагается устанавливать в пределах не более чем одного текста (в узком смысле), так как в этом случае можно исходить из наличия единой когнитивной базы, в пределах которой допустимо говорить о кореферентности.

На этапе поверхностно-синтаксического анализа, не предполагающего обращения к глубинной структуре предложения или тезаурусу [2,3], задача состоит в том, чтобы построить это отношение для случаев, когда кореферентность манифестируется доступным поверхностно-синтаксическому анализу способом. Не предполагается возможность текстуальной взаимозаменяемости имен [ср. 1].

На этом уровне анализа не решается задача собственно определения экстенционала. Так, для текстов в перволичной форме эта задача лежит за рамками синтаксического анализа: например, **я** или **мы** даже в научной статье может означать автора или же автора совместно с читателем, при этом экстенционал политесной формы **мы** отнюдь не всегда - группа исследователей, а в художественном тексте денотатом **я** или **мы** (не говоря о прямой речи) может быть и автор, и кто-то из действующих лиц (если, например, имеет место перволичная форма повествования), и - для **мы** (*мы хорошо знаем*) - некоторый круг - от узкого профессионального до мирового сообщества.

2. Классы рассматриваемых имен.

В линейной структуре текста отношение кореференции может возникать как между полнозначными именами, так и между полнозначными именами и именами дейктическими или с дейктическими компонентами, пополняющими

информацию об экстенционале через посредство связей дейктических элементов с полнзначными.

С точки зрения кореференции имен как референциального тождества могут быть рассмотрены три класса именных выражений, вступающих в отношение референции, для которых, соответственно, и может встать вопрос о кореферентности.

Нас будет занимать кореференция имен, среди которых принято различать

1. дейктические (я, ты, оно, это..., этот, свой, все, тот,...)
2. имена собственные (Венеция, Наполеон, Попов... «Московские новости» («МН»), «Война и мир», «Убить пересмешника», ...)
3. именные группы - дескрипции - выражения с вершиной - именем нарицательным (этот стол, король Франции...)

Обсуждается проблема кореферентности имен этих типов в пределах одного предложения.

Ниже представлены некоторые наблюдения, которые предполагается использовать на этапе анализа кореференции, следующем после сегментации предложения и анализа связей внутри сегмента и между сегментами в системе поверхностно-синтаксического анализа [2,3].

Дешифровочный характер задачи, как и на любом уровне анализа, предполагает возможность неоднозначности решения [2,3].

3. Способы манифестации кореференции на поверхностно-синтаксическом уровне

К синтаксически выраженным относятся отношения, которые I. манифестируются синтагматически, т.е. посредством подчинительной связи имен в графе предложения и II. могут быть заданы линейным порядком, позициями в графе предложения, соотношением поверхностно-синтаксических функций имен, вступающих в отношение кореференции, в графе и условиями конгруэнтности [5].

Первый тип кореференции позволяет устанавливать кореференцию имен всех трех перечисленных типов (дейктических, собственных и дескрипций). Ко второму типу (II) относятся анафорические связи - связи между именами 2-ого и 3-его типа и именами 1-ого - дейктическими.

Кореферентность имен будем условно записывать как равенство вершин именных групп: Имя1 \approx Имя2. Именные группы, если не требуется уточнения, будем называть *именами*.

4. Случаи изоморфизма подчинительной связи и отношения кореферентности.

К кореферентности, выраженной синтагматически, относятся случаи, когда отношение кореференции манифестируется посредством отношения прямого подчинения, т.е. когда это отношение изоморфно синтагматической связи. Обсуждение этого «тривиального» типа представляет интерес с точки зрения того, какие именно синтагматические связи можно использовать как имплицитные кореференцию.

Наиболее вероятны следующие случаи такого изоморфизма.

I. [Имя1 - подлежащее] \approx [имя2 - именная компонента именного сказуемого]. Под именем понимаем любое существительное, полнзначное или местоименное, с зависимыми словами или без таковых. *“Кавказский пленник” - первый неудачный опыт характера, с которым я насилу сладил... Рейнсдорф был человек очень глупый.* (Здесь и далее, если не оговорено другое, приводятся примеры из “Записных книжек” А.С.Пушкина).

II. [Имя1 -] \approx [имя2 - обособленное приложения к имени1(часто имя собственное)]. *Альманах “Северные цветы” разделяется на...; Царь Иван Васильевич во время осады...*

IIIa) [Имя1 - хозяин обособленного приложения] \approx [имя2 - вершина обособленного приложения к имени1(может быть именем собственным)]. *Слышал от старухи, сестры ее - живущей милостынею. Дед мой, Осип Абрамович...; служил во флоте и женился на Марье Алексеевне Пушкиной, дочери тамбовского воеводы, родного брата деду отца моего...; ... князь Кантемир, один из воспитанников Петра...; Кудрявцев, столетний старик, на носилках вынесен был...*

IIIb) [Имя1- слуга предлога] \approx [имя2 - слуга предлога в уточняющем обороте к предложной группе с именем1]

Жена его вчера ночью уехала к своему брату, к дивизионному генералу.

IIIc) [Имя1- хозяин оборота с *как*] ≈ [Имя2 - слуга имени1 в обороте с *как* в значении 'в качестве'] (*Крылов, как сын капитанский, был предводителем...*);

IIIд) [Имя1 - хозяин уточняющего оборота] ≈ [Имя2 в уточняющем обороте, вводимом словосочетаниями *то есть (т.е.), а именно, в частности...*] *Ришелье установил комиссаров, т.е. сановников, временно уполномоченных королем.*

IV. [Имя1- хозяин присубстантивного придаточного] ≈ [Имя2 – местоимение - подчинительный союз в присубстантивном придаточном] (*Вчера был парад, который как-то не удался. "Кавказский пленник" - первый неудачный опыт характера, с которым я насилу сладил...*) В последнем примере - неоднозначность в силу неразрешимости на этом уровне анализа неоднозначности при определении хозяина присубстантивного придаточного: *опыт ≈ который* или *характер ≈ который*)

5. Анафорические связи.

Проблема анафорических связей интенсивно разрабатывалась с точки зрения типологической, структуры высказывания, в рамках порождающих грамматик (ПГ).

В целом эта задача, обсуждаемая обычно с точки зрения теории высказывания или ПГ, для анализа предложения и тем более текста на русском языке в настоящее время не решена.

“Анафорическая связь - это отношение, возникающее между словами во фразе или в тексте, когда смысл одного слова или выражения содержит отсылку к другому при отсутствии синтаксической связи между ними” [6].

Это верно, если под синтаксической связью понимать только непосредственное подчинение и сочинение. Можно считать, что связь синтаксически выражена, если мы на основании линейного порядка слов, информации о сочинении и подчинении в графе и конгруэнтности [5] можем сформулировать условия, позволяющие связь установить. При этом анафорические связи во многих случаях оказываются не менее регулярны (с точностью до синтаксической омонимии), чем синтагматические или сочинительные.

5.1. Дейктические выражения.

Наиболее частые представители класса дейктических элементов текста - местоимения:

1. Возвратные местоимения: *себе,...*; *своей...*
2. Выражение *друг друга*.
3. Личные местоимения 3-его лица существительные и прилагательные: *он, она, оно, они; его, ее, их*.
4. Личные местоимения 1-ого и 2-ого лица и прилагательные от них: *я, мы,...*; *мой, наш, твой, ваш...*
5. Относительное местоимение: *который...*
6. Подчинительные союзы и местоимения - подчинительные союзы, которые могут выступать в функции “связок” в присубстантивных придаточных: *кто, что, где, когда...*
7. Указательные местоимения *этот, тот*.

5 и 6 классы представляют дейктические слова, кореферентность которых антецеденту - полнозначному или дейктическому - изоморфна отношению непосредственного подчинения. Таким образом, в этом случае задача решается уже в процессе анализа межсегментных связей.

5.2. Теория связывания.

Некоторые наблюдения, полезные для решения проблемы определения антецедента анафорического замещения, находим в теории связывания, разрабатываемой в порождающих грамматиках (ПГ). Понятие связывания заимствовано из логики: это функция оперетора, превращающая сентенциальную форму в полное высказывание.

“ИГ α называется связанной антецедентом β , если β командует α ... и обе эти группы помечены одним референциальным индексом”.

В ПГ по отношению к связыванию все выражения делятся на анафоры, прономиналы и референциальные выражения. Анафоры (в частности, возвратные и взаимные местоимения) должны быть связаны (иметь антецедент) в пределах некоторой составляющей, в которую они входят (“непрозрачной области”, которая и является для

анафоров областью связывания). Пронаминалы, напротив, не могут быть связаны в пределах некоторой составляющей (“непрозрачной области”, которая не м.б. их областью связывания). Референциальные выражения не могут быть связаны ни в какой структурно определяемой области, в которую они входят [9].

Условия связывания относятся к базису синтеза и предполагают препозициональное знание о кореферентности α и β . При анализе встает обратная - дешифровочная задача. В нашем случае - на основании отношений подчинения, сочинения и линейного порядка слов в предложении.

Безусловно необходимыми при установлении кореферентности являются условия конгруэнтности местоименных существительных и полнозначных имен, входящие в один ряд с согласованием и управлением, - согласование по числу и роду [5]. Различия, связанные с синтаксическими ролевыми различиями действительных прилагательного и существительного здесь не рассматриваются. Краткие комментарии к ситуациям – попытка задать круг включаемых в рассмотрение явлений - не претендуют на строгость формулировок.

5.2. Антецедент возвратных местоимений.

Возвратные местоимения (Pron^{возвр}) - наиболее изученный класс слов, вступающих в анафорические отношения [5,8,9] К этому классу относятся 1. Pron^{возвр} - существительные.: *себе*^{Дат.}, *себя*^{Род., Вин.}, *собой*^{Твор.}, *себе*^{Предл.}; 2. Pron^{возвр} - прилагательное *свой*; 3. *сам* (-а, -о, -и) как отдельное слово или в словосочетании с (+ факультативно предлог) + *себе*^{Дат.}, *себя*^{Род., Вин.}, *собой*^{Твор.}, *себе*^{Предл.} (и наречные словосочетания *сами собой*, *сам по себе*, *сама для себя* и т.д.)

Pron^{возвр} появляется в именной группе дополнения и при этом замещает в дополнении или определении к дополнению субъект предиката. Соответственно, антецедент Pron^{возвр} - субъект предиката.

В поверхностно-синтаксической структуре в русском языке возможны 3 способа представления этой ситуации: внутри одного сегмента с одним предикатом (**Ситуация 1**) и между именами, находящимися в разных сегментах (**Ситуации 2 и 3**) внутри одного сегмента с вложенным в его структуру предикативным именем (**Ситуация 4**).

Ситуация 1. Pron^{возвр} (существительное или прилагательное) в сегменте - простом предложении (главном или придаточном). Его антецедент - имя1= подлежащее.

Каждый владелец управлял в своем участке по-своему, устанавливал свои законы.. и старался окружить себя достаточным числом приверженцев для удержания и повиновения своих вассалов...; Ермолов принял меня с обыкновенною своей любезностью. Другие литераторы позволили себе посмеяться над нетерпимостью дворян-журналистов.

В случаях, когда субъект выражен не именительным падежом (например, именем в Дат.п.), Pron^{возвр} кореферентно этому имени (имени в Дат.п.: *Отцу неприятно (приятно, обидно, странно, больно, забавно, противно, хочется и т.д.) считать себя обиженным*).

Pron^{возвр} и его антецедент могут оказаться в разных сегментах в двух случаях: **Ситуация 2** и **Ситуация 3**.

Ситуация 2. Pron^{возвр} - дополнение или определение к дополнению предикатива - вершины деепричастного оборота. Это возможно в силу того, что субъект деепричастия совпадает с субъектом главного по отношению к нему предложения, предикату которого подчинен деепричастный оборот: поверхностно-синтаксическое подчинение деепричастия предикату является по сути тождественным сочинению предикатов.

Каждый владелец управлял в своем участке..., соблюдая свои выгоды,...; Екатерина явно знала духовенство, жертвуя тем своему неограниченному властолюбию...; Налив рюмку себе, велел он и мне поднести...

Ситуация 3. Pron^{возвр} = дополнение или определение к дополнению в обособленном или необособленном согласованном определении с вершиной A^v, то антецедент - хозяин согласованного определения. В этом случае субъектом предикативной вершины обособленного оборота является хозяин этого оборота и, соответственно хозяин и является искомым антецедентом.

Даже люди, выдающие себя за усерднейших почитателей прекрасного пола,...; Во Франции, в земле, прославленной своей учтивостью...; Они относят к нему все произведения, носящие на себе печать уныния или мечтательности.

Ситуация 4. Pron^{возвр} – в именной группе, подчиненной предикативному имени. Ситуация, порождающая неоднозначность.

Она должна в нем ненавидеть убийцу брата своего... (Пример из [7]). В подобных случаях порождается неоднозначность в силу вложенности в предикативную структуру простого предложения предикативного имени (*убийца*), являющегося субъектом предикативной ситуации, объект которой – *брат*.

5.3. Антецедент личного местоимения.

5.3.1. Анафорическое замещение в ситуациях сочинения или соподчинения.

Одним из самых распространенных случаев замещения имени личным местоимением - случай элиминирования потенциального повтора в ситуации некоторого синтаксического “параллелизма” фрагментов структуры предложения - при сочинении имен, сочинении как эллипсисе, при соподчинении, при сочинении предикатов и сочинительном эллипсисе.

Ситуация 1. Pron^{личн} в сочиненных или соподчиненных группах.

Если есть несколько непосредственно сочиненных или соподчиненных имен, и Pron^{личн} – в не первой по порядку следования в линейной структуре S группе, его антецедент – конгруэнтное ему имя, предшествующее ему и сочиненное с ним или с его хозяином.

Он и брат его...; Простительно выходишь не любить ни России, ни истории ее, ни славы ее. ...велел он расстрелять в Берде Харлову и 7-летнего брата ее.

5.3.2. Ситуации с сочинением предикатов или эквивалентным ему подчинением:

замещение при “параллельном” подчинении имен сочиненным предикатам или в ситуации поверхностно-синтаксических эквивалентов сочинения, когда у предикатов один субъект, или в сочиненных простых предложениях с одинаковыми или экстенционально тождественными субъектами.

Ситуация 2. Сочинение предикатов. Антецедент Pron^{личн} - конгруэнтное имя в линейно предшествующей ветке сочиненного предиката.

Но они гнушаются просторечием и заменяют его простомыслием.

Здесь нашел я у коменданта рукопись «Кавказского пленника» и, признаюсь, перечел его с удовольствием. Вигель получил звезду и очень ею доволен.

Один из гостей взял ружье покойника, сдул порох с полки и положил его подле тела. Последний пример позволяет неоднозначную интерпретацию, так как сочинены три предиката, два из которых - *взял ружье* и *сдул порох* - имеют полнзначные дополнения, возможны, соответственно, два варианта “дешифровки”: *положил его ≈ ружье* | *положил его ≈ порох*.

Ситуация 3. Pron^{личн} в ситуации сочинения инфинитивов, когда

дополнение или определение при дополнении в ситуации сочинения инфинитивов (сочинение инфинитивов здесь - эллипсис во избежание повтора при сочинении одинаковых глаголов с подчиненными им разными инфинитивами)

Он предпочел впречь целое стадо в огромную бричку, нагруженную запасами всякого рода, и торжественно перевезти ее через снеговой хребет; мы хотели бы обратиться назад и взглянуть... на ее старинные памятники, сравнить их с этой бездной романсов.

Ситуация 4. Pron^{личн} в ситуации сочинения причастных оборотов. Pron^{личн} кореферентно конгруэнтному имени в дополнении к сочиненному причастию.

... выдумал сказку о богородице, будто бы явившейся к умирающей матери и приказавшей ей надеяться...;

Ситуация 5. Причастный оборот при субъекте - аналог сочинения предикатов. Pron^{личн} – в дополнении в сегменте – простом предложении (м. б. придаточным) или в причастном обороте, хозяин которого – субъект простого (придаточного). *Некто Карцев, женатый на парижской девке в 1814 году, развелся с нею...*

Ситуация 6. Pron^{личн} - дополнение в сочиненных простых (главных или придаточных) с одинаковыми или кореферентными субъектами (или субъект в эллипсисе) Антецедент - имя в дополнении к сочиненному предикату.

Я никогда не хлопотал о счастье, я мог обойтись без него. Совпадающие по субъекту сочиненные простые (м.б. в составе придаточных) интерпретируем аналогично сочинению предикатов, т.е от сочиненного предиката вниз до подходящего по условиям конгруэнтности имени:

Процедура поиска антецедента - восходящее движение по графу до предиката:

него ← без ← обойтись ← мог ↔^{Соч} не хлопотал → о → **счастью**.

Ситуация 7. Присубстантивное придаточное с подлежащим, кореферентным подлежащему главного, т.е. случай, аналогичный сочинению предикатов с одним субъектом. Pron^{личн} - дополнение при глаголе в главном или присубстантивном придаточном с подлежащим, кореферентным субъекту главного. Антецедент (в зависимости от линейного порядка) – конгруэнтное имя в дополнении к предикату в главном или присубстантивном.

Завистник, который мог освистать Дон Жуана, мог отравить его творца. Заметим, что при сегментации этого предложения возникает неоднозначность: отрезок со вторым глаголом *мог* можно понять как 1. сочиненный с предикатом присубстантивного придаточного при простом-назывном *Завистник* или 2. как часть простого-главного *Завистник мог отравить его творца* и присубстантивное придаточное *завистник, который мог освистать Дон Жуана*. [3]. Интересно, что в нашем случае варианты сегментации в силу кореферентности подлежащих не порождают неоднозначности анафорической связи.

Процедура поиска антецедента – восхождение по связям до сочинения предиката (здесь – личной формы глагола) и

от сочиненного вниз до подходящего по условиям конгруэнтности имени:

его ← *творца* ← *отравить* ← *мог* ↔^{Соч} *мог* → *освистать* → *Дон Жуана*

Ситуация 8. Как уже отмечалось, отношение подчинения деепричастия - вершины деепричастного оборота предикату является по сути поверхностно-синтаксическим аналогом сочинения предикатов. Аналогично сочинению предикатов устанавливается и отношение кореференции (Ситуация 2.1).

Казачка Разина...пригребала палкою к себе мимо плывущие трупы, переворачивая их и приговаривая...;...и предал этерию, представляя ее Императору Александру отраслю карбонаризма. Миних спас Ганнибала, отправя его тайно в ревельскую деревню...; Выкупив бриллианты Натальи Николаевны,...я принужден был их перезаложить...;

Приступая к изучению нашей словесности, мы хотели бы обратиться назад и взглянуть...на ее старинные памятники, сравнить их с... Посадив на стул одного майора, они подходили к нему с шутками. В последнем примере у первого предиката два полнозначных слуги (*посадили на стул, посадили майора*, что порождает неоднозначность: *подходили к стулу \ к майору*).

Процедура поиска antecedента – восхождение по графу до уровня предиката и затем - движение вниз до подходящего (с учетом вариантов) по условиям конгруэнтности имени:

нему ← к ← *подходили* ↔ *посадив* → на → *стул*

нему ← к ← *подходили* ↔ *посадив* → *майора*

5.3.3. Анафорическое замещение при ролевой инверсии.

В поверхностно-синтаксической структуре предложения анафорическое замещение часто встречается в ситуации, которую можно назвать “зеркальной ролевой инверсии”, когда имя, исполняющее в главном сегменте, в котором находится “хозяин” подчиненного, роль дополнения, становится субъектом подчиненного, а субъект главного - дополнением в подчиненном.

Ситуация 9. R^{пичн} в придаточном, определяющем имя в дополнении к предикату: Antecedent R^{пичн} - субъект главного (при условии их конгруэнтности).

А публика...всегда соглашается с тем, кто последний жалуется ей. (тем ≈ кто; публика ≈ ей). **Мы** так привыкли читать *ребяческие критики*, что они даже нас не смешат.

Ситуация 10. R^{пичн} в причастном обороте, определяющем имя в дополнении к предикату. Antecedent R^{пичн} - субъект предиката, - хозяина дополнения, определяемого оборотом.

Кудрявцев ..., на носилках вынесен был в церковь, близ его загородного дома находившуюся.

В некотором Азиатском народе мужчины каждый день, восстав ото сна, благодарят Бога, создавшего их не женецинами. Мать его жила в белорусской деревне, пожалованной ей Государыней. Стесненный Терек с ужасным ревом бросал свои аспидные волны через камни, преграждающие ему путь. Как и в Ситуации 5 имеет место инверсия, которая заключается в том, что камни одновременно дополнение предиката в главном и субъект предиката, выраженного причастием.

Процедура поиска antecedента: вверх от R^{пичн} по графу связей к подлежащему. Если представлять связь (предикат R субъект) как двунаправленную, можно не оговаривать перемены направления движения по графу, и путь поиска antecedента можно представить как прямое восхождение по графу синтагматических связей: *ей* ← *жалуется* ← *кто* ← *с тем* ← *соглашается* ↔ *публик*; *ему* ← *преграждающие* ← *камни* ← *через* ← *бросал* ↔ *Терек*.

Ситуация 11. Похожая ситуация возникает в придаточном, замещающем некоторую валентность предиката главного, в этом случае инверсии в главном сегменте нетдополнения, определяемого придаточным, т.е. происходит “стяжение”: дополнение - само придаточное, и субъект главного выступает в нем в роли дополнения (часто - объекта): *...если же крестьяне узнают, что правительство или помещики намерены их кормить...*

Троцкий не может понять, что с ним делается.

6. Заключение.

Описанные ситуации не исчерпывают, естественно, всех возможных случаев описываемого явления, однако представляют значительную их долю.

Следующим объектом штудий необходимо должно явиться 1. анафорическое замещение при сочинении простых предложений и 2. анафорическое замещение в цепочке отдельных предложений текста.

Результаты предполагается использовать в разрабатываемой в настоящее время системе поверхностно-синтаксического анализа русского текста с последующим анализом кореференции. [2,3].

Литература

1. Н.Д.Арутюнова. Предложение и его смысл. М. Наука 1976.
2. Т.Ю.Кобзарева, Д.Г.Лахути, И.М.Ножов. Модель сегментации русского предложения. Диалог'2001. т.2. Аксаково.2001.
3. Т.Ю.Кобзарева. Некоторые аспекты анализа сочинения при сегментации русского предложения. Труды КИИ'2002. М. 2002.
4. М.А.Кронгауз. Семантика. М.2001.
5. И.А.Мельчук. Согласование, управление, конгруэнтность. ВЯ 1993 Наука №5 с. 16-58
6. Е. В.Падучева. Высказывание и его соотносительность с действительностью. (Референциальные аспекты семантики местоимений). М. Наука. 1985.
7. Н.И.Перцов. О неоднозначности в поэтическом языке. ВЯ №3 2000 с. 55-82.
8. И.И.Ревзин. Структура языка как моделирующей системы. М. Наука. 1978.
9. Я.Г.Тестелец. Введение в общий синтаксис. М. 2001.