Разметка кореференции на синтаксически аннотированном корпусе чешских текстов[1]

Coreference annotation in Prague dependency treebank

Недолужко А. (nedoluzko@ufal.mff.cuni.cz)

Карлов университет, Прага, Чехия

В докладе представлена схема разметки кореференции на синтаксически аннотированном корпусе чешских текстов PTD. Рассматриваются три этапа разметки — разметка грамматической кореференции, где антецедент высчитывается на основе грамматических правил данного языка, разметка прономинальной текстовой кореференции и расширенная схема разметки именной текстовой кореференции и ассоциативной анафоры. Разметка грамматической и прономинальной кореференции была проделана на всем корпусе PTD, разметкой именной кореференции и ассоциативной анафоры занимается автор данного доклада в настоящее время. В докладе рассматриваются некоторые трудности классификации примеров, приводятся первые результаты.

1. Общие сведения

Синтаксически аннотированный корпус чешского языка (PDT) — это проект лингвистической разметки текстов, разрабатываемый в Институте формальной и прикладной лингвистики физико-математического факультета Карлова университета в Праге. Разметка проводится частично автоматически на трех уровнях — морфологическом, поверхностно-синтаксическом и глубинно-синтаксическом (подробнее с проектом можно ознакомиться Hajičová 2006, Недолужко 2007). В данном докладе речь пойдет о разме тке кореференции, реализуемой вручную и частично автоматически на глубинно-синтаксическом уровне.

В настоящее время аннотирование кореференции проводится с различной степенью подробности в большинстве синтаксически размеченных корпусов. Прономинальная кореференция представлена в американском PennTreebank (http://www.cis.upenn.edu/~treebank), концепции разметки именной кореференции представлены в проектах MUC-7 (Hirschman, 1998), MATE (Poesio, 2004), DRAMA (Passonneau, 1997), PoCoS (Chiarcos, Krasavina 2005), аннотация ассоциативной анафоры проводится в рамках проектов GNOME (на основе MATE), DRAMA, планируется в PoCoS и т. д.

В аннотации PDT 2.0 кореференция делится на грамматическую и текстовую. Кроме того, аннотируется т. наз. ассоциативная анафора (bridging) и некоторые особые случаи (экзофорическая отсылка и отсылка к большему, чем одно предложение, сегменту текста). Для аннотирования кореференции используется `id` антецедента, к которому отсылает `id` узла анафоры. Разметка кореференции приводилась в три этапа. Первый этап — разметка грамматической кореференции (см. 2), второй этап — разметка т.наз. текстовой прономинальной кореференции (см. 3), третий этап состоит из разметки именной кореференции и ассоциативной анафоры (см. 4). Далее будут представлены эти три этапа с особым акцентом на последний, которым автор доклада занимается в настоящее время.

2. Разметка грамматической кореференции

В случае грамматической кореференции антецедент высчитывается на основе грамматических правил языка. Грамматическая кореференция практически никогда не переходит границ предложения, ее всегда можно представить как отсылку одного узла к другому, следовательно ее аннотирование легко автоматизируется. К грамматической кореференции относится:

[1] эта работа была поддержана грантом GACR 405/09/0729

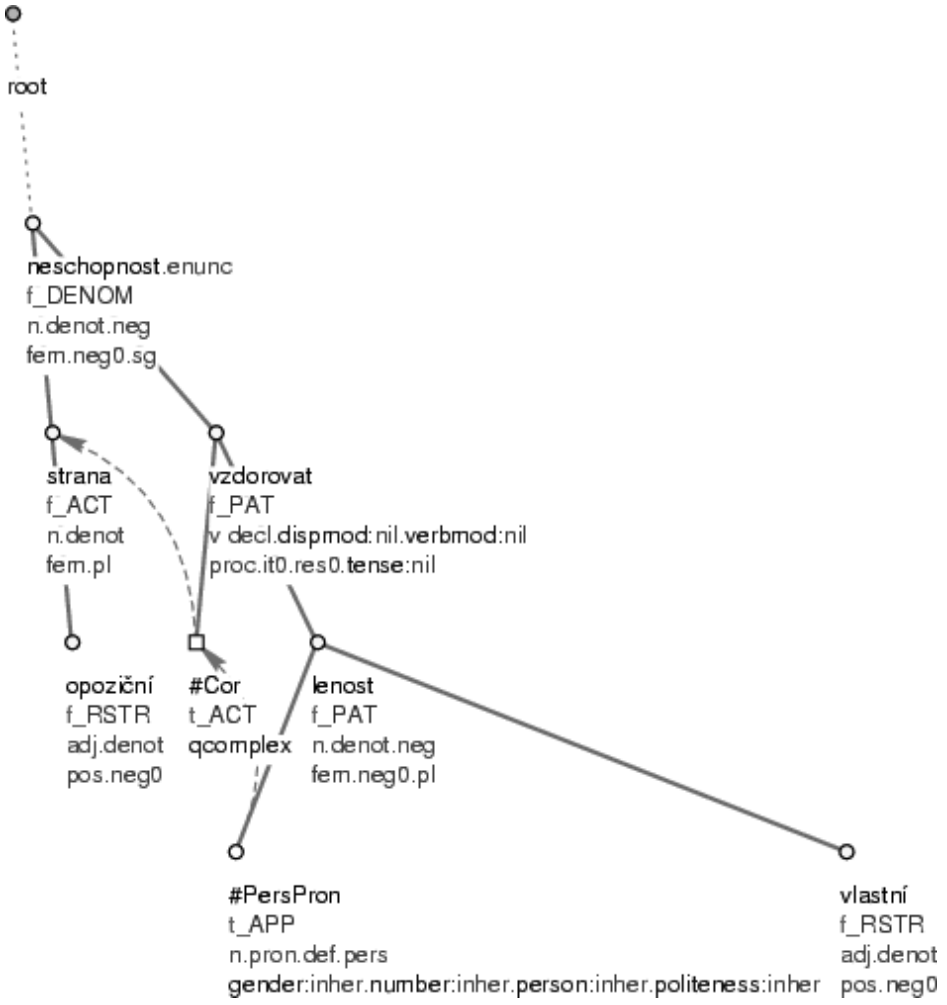

Рис. 1. Грамматическая кореференция

(1) *Neschopnost opozičních stran vzdorovat své vlastní lenosti.* — Неспособность оппозиционных партий бороться с собственной (со своей) ленью.

- кореференция возвратных местоимений, в случае, если они являются самостоятельным членом предложения (возвратное *se* («*ся*»)[2], лексемы *sebe* (*себя)* и *svůj* (*свой)*). Все возвратные местоимения имеют общую лемму *se* («ся») и отсылают к субъекту предложения, к ближайшему узлу с функтором ACT (агенс) — первично к агенсу той же клаузы, в случае, если он там отсутствует — к агенсу главного предложения. (см. рис. 1)
- кореференция относительных средств. К ним относятся относительные местоимения и наречия, относительные придаточные предложения и т. д. Ср. *člověk, který pije (человек, который пьет)*; *ve městě, kde se mi tak líbilo* (*в городе, где мне так понравилось*) и др.). В глубинно-синтаксическом дереве стрелка грамматической кореференции ведет от относительного местоимения *(который, где)* к управляющей именной группе (соответственно *человек, город*).
- кореференция в т. наз. контролирующих конструкциях (у некоторых глаголов, заданных списком в документации по разметке глубинно-синаксического уровня (Mikulová, 2005), напр. *стесняться, забыть, хотеть, научить* и др., один из актантов которых обязательно кореферентен с определенным актантом зависимого от них глагола в инфинитиве — напр. *zapomenout přečíst* (*забыть прочитать*)). При восстановлении модели управления зависимого глагола, его невыраженный кореферентный актант имеет лемму #Cor, от которого ведет стрелка грамматической кореференции к соответствующему актанту управляющего глагола. (см. рис. 1)
- кореференция актантов в реципрокальных конструкциях. Один из актантов имеет восстановленную лемму #Rcp, откуда ведет стрелка грамматической кореференции на лексически выраженный кореферентный актант (см. рис. 2).

[2] В чешском языке возвратное местоимение «ся» всегда является отдельной лексемой (клитикой).

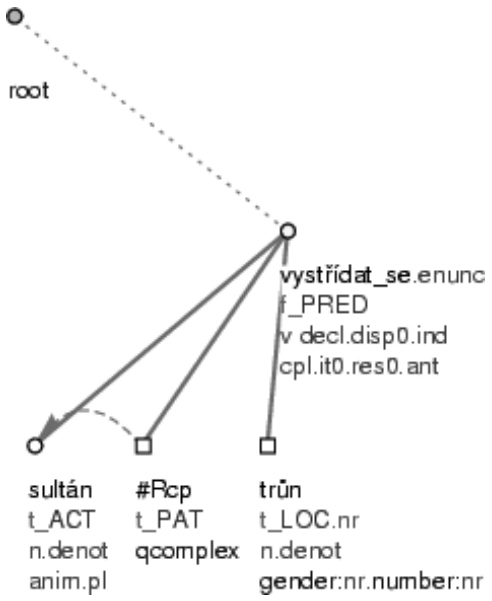

Рис. 2. Грамматическая кореференция

(2) *Sultáni se vystřídali na trůnu.* — Султаны поменялись местами (поменялись) на троне

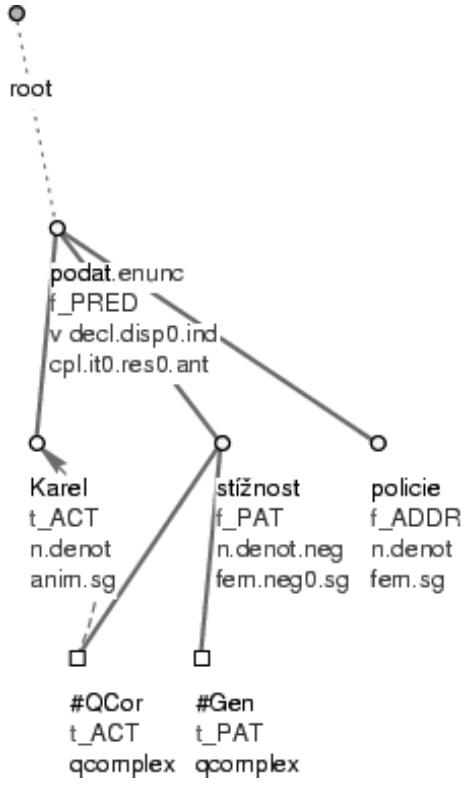

Рис. 3. Грамматическая кореференция

(3) *Karel podal stížnost policii.* — Карл подал жалобу в милицию.

- кореференция в т.наз. квазиконтролирующих конструкциях (в случае составного предиката, именной частью которого является имя существительное, имеющее модель управления, напр. *подать жалобу в милицию*). При восстановлении модели управления зависимого существительного, его невыраженный агенс имеет лемму #QCor, от которого ведет стрелка грамматической кореференции к агенсу управляющего глагола. (см. рис. 3)
- кореференция у дополнений с двойной зависимостью, выраженных формой глагола. Отношением кореференции связан восстановленный актант дополнения, выраженного формой глагола (причастием, деепричастием или инфинитивом) с актаном управляющего предиката. (см. рис. 4)

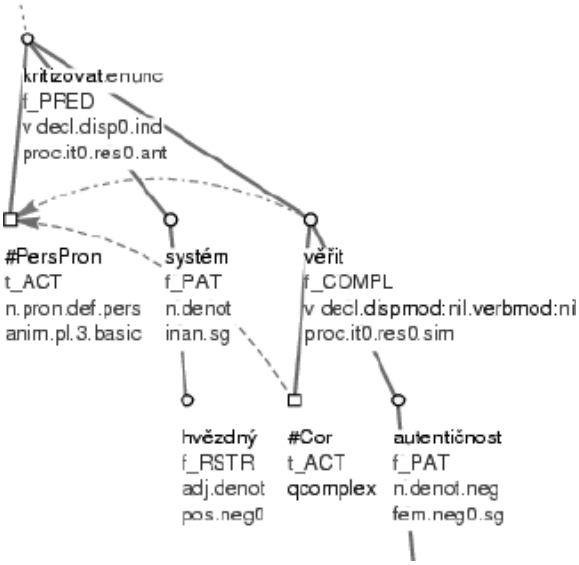

Рис. 4. Грамматическая кореференция

(4) *Kritizovali hvězdný systém, věříce v autentičnost…* — Они критиковали звездную систему, веря в истинность…

3. Разметка прономинальной текстовой кореференции

Текстовая кореференция понимается как использование различных языковых средств для анафорической (реже катафорической) отсылки. Эта отсылка реализуется не только за счет грамматических средств языка, но и на основании знания контекста. Текстовая кореференция может легко переходить границы предложения. Разметка текстовой кореференции проводилась вручную на всем корпусе текстов PDT. Текстовая прономинальная кореференция размечена в PDT 2.0 в следующих случаях:

- в качестве анафора выступают личные и притяжательные местоимения третьего лица. Кореференция местоимений первого и второго лица не размечается. Местоимения (в том числе эл-

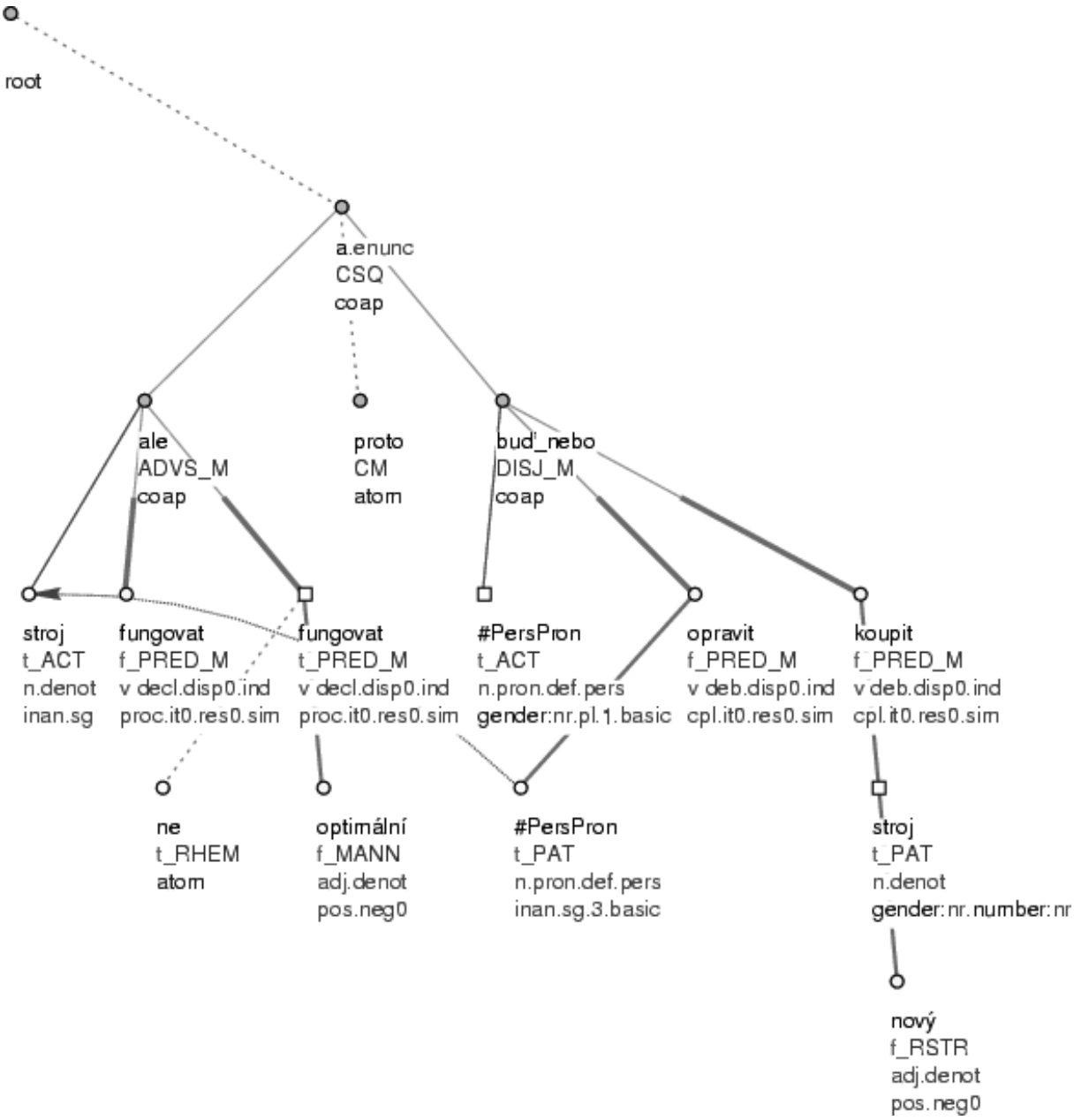

Рис. 5. Текстовая прономинальная кореференция

(5) *Stroj funguje, ale ne optimálně, a proto ho musíme buď opravit, nebo koupit nový.* — Прибор работает, но не оптимально, поэтому его нужно либо починить, либо купить новый.

липтированные) на глубинно-синтаксическом уровне имеют лемму `#PersPron` (см. рис. 5).

- в качестве анафора выступает указательное местоимение *этот* в субстантивной функции.
- в качестве анафора выступает эллиптированное и восстановленное на глубинно-синтаксическом уровне местоимение 3-го лица. Являясь языком pro-drop, чешский язык имеет сильную тенденцию опускать личные местоимения в анафорических конструкциях (ср. *cz О Nechtěl to říkat* vs. rus. *Он не хотел этого говорить*). На глубинно-синтаксическом уровне в PDT эти местоимения воостанавливаются, и им присваивается тектограмматическая лемма `#PersPron`. Информация о (не)выраженности этой леммы на поверхностном уровне содержится в атрибуте `is_generated`.

3.1. Отсылка к сегменту текста

Отсылка к сегменту текста имеет место в случае, когда либо антецедент местоимения состоит из более чем одного предложения, либо высчитывается на основании общего контекста. Информация об отсылке к сегменту текста фиксируется значением `segm` атрибута `coref_special`.

3.2. Дейксис

Отсылка к объектам внеязыковой действительности обозначается значением `exoph` атрибута `coref_special`.

4. Разметка именной текстовой кореференции и ассоциативной анафоры

4.1. Разметка именной текстовой кореференции

На данном этапе размечаются референциальные цепочки, где в качестве анафора выступают в основном имена существительные и некоторые наречия (*там*, *тогда* и др.). В некоторых случаях в отношении кореферентности могут участвовать прилагательные (притяжательные прилагательные и прилагательные, образованные от имен собственных) и числительные (выступающих в субстантивной функции и релевантных для связности текста). Технически разметка именной текстовой кореференции является частью предшествующей ей разметки прономинальной кореференции (используется `id` антецедента, к которому отсылает `id` узла анафоры, атрибут `coref_text.rf` содержит `id` кореферентного узла), однако добавляется информация о типе кореферентного отношения (атрибут `informal-type`). Отношение текстовой кореференции не фиксируется между субъектом и именной частью составного именного сказуемого, а также между узлами, находящимися в отношении аппозиции. Идентичность их референтов следует из синтаксической структуры дерева зависимостей.

При разметке именной текстовой кореференции используется 4 типа отношений:

- дефолтный тип 0 (значение `0` атрибута `informal-type`). Отношение между конкретнореферентными ИГ, причем анафор не является гиперонимом или синонимом ИГ антецедента. К этому типу относятся повторы ИГ антецедента (*женщина — женщина*), повторы ИГ антецедента с идентификатором (*женщина — эта женщина*), ИГ с существительным, антецедентом которого является местоимение или эллипсис, являющиеся звеном цепи прономинальной кореференции (таким образом достраиваются цепочки прономинальной кореференции, ср. женщина — она — женщина), частичные повторы ИГ антецедента (общество — акционерное общество) и др.
- синонимия в широком смысле (значение `SYN` атрибута `informal-type`). Обозначается, если анафорический член и ИГ антецедента — различные номинации. Помимо действительной синонимии, к этой группе относятся напр. такие случаи, как имя собственное — имя нарицательное (*Петя — раздолбай*), сокращение — полное название (*НДС — налог на добавленную стоимость*) и др.
- гиперонимия (значение ER атрибута `informal-type`). Этот тип не совсем соответствует своему названию, т.к. в процессе аннотирования его наиболее типичные пары (*яблоко — фрукт*) в результате нечеткой границы с предыдущим типом перешли в тип `SYN`. На настоящий момент тип `ER` приписывается в основном отсылкам на ситуацию (*Начальник заставил нас приходить вовремя. Это решение никому не понравилось*) и в случае т. наз. автонимной анафоры (отношения между ИГ *Адольф Гитлер — это имя, радуга — это слово* и т. д.)
- кореференция нереферентных и родовых ИГ (значение `NR` атрибута `informal-type`). Этот тип несколько проблематичен, т.к. решение связывать кореференцией ИГ, которые не обладают конкретной референцией, не является полностью интуитивным. Тем не менее зачастую нереферентные ИГ способны вступать в анафорические отношения наравне с референтными, в том числе являться антецедентами местоимений (Падучева 1985), поэтому не могут быть исключены из кореферентных цепочек. Пример пары кореферентных ИГ типа NR в (6):

(6) *Paláce neznamenají přepych. Ač se to na první pohled nezdá , obývání klasických renesančních a barokních palácůs velikými , řetězovitě propojenými místnostmi není žádné terno.* — Дворец не значит роскошь. На первый взгляд так не кажется, но обитание в о дворцах {`coref_text`, тип `NR` на «дворец»} в стиле барокко или ренессанса с огромными комнатами, расположенными анфиладой, не так уж безоблачно прекрасно.

Проблематичным является тот факт, что в произвольном корпусе текстов встречается большое количество нереферентных ИГ, отсылающих в принципе к одному и тому же, но не вступающих между собой в анафорические отношения. На данный момент мы не можем предложить алгоритм проведения четкой границы между нереферены (родовыми) ИГ, кореферентность которых является релевантной для связности текста, и просто повторяющимися ИГ с родовым статусом и отсылающими к одному и тому же, поэтому мы отдаем предпочтение аннотации кореферентности перед наличием анафорического отношения и связываем такие ИГ текстовой кореференцией с типом `NR`. Проблематичным также часто оказывается вопрос

о кореферентности ИГ с неконкретнореферентным денотативным статусом — при вторичном просмотре пар с отмеченной кореферентностью этого типа находится множество примеров, где кореференция не должна была бы быть обозначена.

Среди нереферентных ИГ не проводится различие на чистый повтор, синонимичные и гиперонимичные номинации. Это различие касается только ИГ с конкретной референцией. См пример (7):

(7) *Na telefonní číslo 855 44 33 bude jistě volat mládež s různými problémy. Doufejme, že linka si časem vydobude mezi dětmi takovou autoritu, aby se na ni obracely i ty, které jsou skutečně ohrožovány.* — По телефонному номеру 855 44 33 молодежь будет звонить с различного типа проблемами. Будем надеяться, что этот номер со временем достигнет такой популярности среди ребят {`coref_text`, тип `NR` на «молодежь»}, что по нему будут звонить и дети, которым действительно что-то угрожает.

Отдельную проблему представляют **абстрактные имена**. Проблематично уже само разделение имен на конкретные и абстрактные (Степанов 2004, Падучева 1986 и др.) Однако даже если предположить, что эта проблема решена, вопрос определения их денотативного статуса остается открытым. В нашей разметке кореференция абстрактных имен обозначается по умолчанию типом `NR`, однако не совсем последовательно. Если ИГ обладает абстрактной семантикой, но при этом очевидно конкретной денотативностью, разметчик вправе обозначить и дефолтный тип `0`. Эта конвенция однако является спорной и находится в стадии обсуждения. Ср. тип `NR` в (8) и тип `0` в (9):

(8) *Tímto faktorem je podnikatel — inovátor , který se snaží o zisk , a proto logicky nemůže existovat ve stavu statiky, která nezná ani zisk , ani ztrátu.* […] *Na konci tohoto procesu se systém vrátí ke statické rovnováze, v níž nebudou opět ani zisky, ani ztráty.* — Этим фактором является предприниматель-инноватор, который пытается получить прибыль и потому не может находиться в статичном состоянии, которому неизвестны ни прибыль {`coref_text`, тип `NR` на «прибыль»}, ни убыток. […] В конце этого процесса система снова возвращается к статическому равновесию, в котором снова не будет ни прибыли {`coref_text`, тип `NR` на «прибыль»}, ни убытка.

(9) *Televize dává příležitosti k podnikání.* […] … *nevyužité možnosti stále má televize zejména při regionálním vysílání.* — Телевидение располагает к предпринимательству. […] … неиспользованными возможностями обладает телевидение {`coref_text`, тип `0` на «телевидение»} прежде всего в региональном вещании.

Похожим образом разрешается разметка кореференции **имен действий**. Имена действий чаще бывают конкретны и соотносимы с реальной ситуацией, однако возникает проблема временной локализации действий и возможности кореференции ИГ расположенных на различных участках временной оси (см Падучева 1986). В данном случае решение о наличии стрелки кореференции часто бывает основано на языковой интуиции разметчика.

При разметке грамматической и текстовой кореференции выдерживается принцип сохранения референциальной цепочки, контролируемый частично автоматически. Если разметчик устанавливает отношение кореферентности с узлом, к которому уже ведет стрелка, новое отношение автоматически устанавливается с последним (самым правым) узлом.

4.2. Разметка ассоциативной анафоры (т. наз. bridging anaphora)

Параллельно с разметкой именной текстовой кореференции проводится разметка т.наз. ассоциативной анафоры (bridging anaphora). Анафорический член и антецедент в данном случае уже не кореферентны, но между ними имеется семантическое отношение определенного типа.

При аннотации PDT действуют некоторые конвенции выбора той или иной связи в сомнительных случаях. Одной из основных конвенций является предпочтение текстовой кореференции перед ассоциативной анафорой.

Наличие разметки ассоциативной анафоры связано с общей структурой дерева зависимостей глубинно-синтаксического уровня PDT. Ассоциативная анафора не аннотируется, например, если узел участника отношения является непосредственным потомком антецедента с определенным функтором (PAT, APP, AUTH и др.[3]), если отношения между участниками отношения уже выражены грамматическим функтором или синтаксической структурой дерева и т.д. (Nedoluzhko 2007)

В отличие от текстовой кореференции, разметка ассоциативной анафоры затрагивает практически только те узлы, которые соответствуют в тексте полнозначным лексемам. Ссылка на эллиптированные местоимения, союзы и знаки препинания возможна только в том случае, если другого не позволяет структура дерева.

- С технической точки зрения разметка ассоциативной анафоры — это отсылка узла анафора к `id` антецедента, информация о связи со-

[3] к описанию значения функторов см (Hajičová и др. 2006)

держится в атрибуте `bridging`. Информация о типе отношения отображается в атрибуте `informal-type`. Аннотация ассоциативной анафоры не является дополнением референциальной цепочки, состоящей из отношений грамматической и текстовой кореферентности, а существует параллельно. Референциальная цепочка ассоциативной анафоры не удерживается (по крайней мере, не удерживается последовательно).

При разметке PDT выделяются и размечаются следующие типы ассоциативной анафоры:

- отношение множество-подмножество/элемент множества (значения `SUB_SET` и `SET_SUB` атрибута `informal`-type в зависимости от направления отношения). Типичные примеры: *мушкетеры — Атос, Портос, Арамис; семинары — первый семинар, последний семинар*. Ср. также (10):

(10) *Na rozdíl od dobře vybaveného FS dnes nikdo z téměř dvou stovek poslanců kromě předsedy a místopředsedů sněmovny nemá svou kancelář, pracovní stůl, židli a telefon.* — В отличие от хорошо оснащенной Федерального парламента, сегодня ни у кого из почти двухсот депутатов, кроме председателя {`bridging`, `typ=SET`, на «poslanec (депутат)»} парламента и зампредседателей {`bridging`, `typ=SET`, на «poslanec (депутат)»} нет своего кабинета, рабочего стола, стула и телефона.

- отношение часть — целое (значения `PART_WHOLE` и `WHOLE_PART` атрибута `informal-type` в зависимости от направления отношения). Типичные примеры: *комната — потолок, рука — палец* и др. Как часть — целое аннотируются также неотделимые части в географических названиях, напр. *ФРГ — Бавария — Мюнхен*. Граница между отношениями «часть — целое» и «множество — подмножество» не всегда является достаточно отчетливой. Во многих случаях решение зависит только от исчисляемости объектов, входящих в данное отношение (напр. *заграница — Германия vs. иностранные государства — Германия; текст — предложение* и др.). Возможно, в дальнейшем эти два типа можно совместить (ср. проекты PoCoS, MATE и др.), но пока мы размечаем их отдельно.
- отношение дискурсивного контраста, имеющего значение для связности текста (значение `CONTRAST` атрибута `informal-type`). Этот тип частично пересекается с размеченным на всем корпусе PDT актуальным членением (Hajičová 2006, коротко также в Недолужко 2008), но не полностью его копирует. Члены отношения ассоциативной анафоры типа `CONTRAST` могут находиться в предложении как в позиции контраста, так и в позициях топика и фокуса; кроме того, ассоциативный контраст не ограничен рамками предложения. Ср. пример (11), где ИГ *коровы* расположена в фокусе:

(11) *Lidi nežvýkají, to jenom krávy*. — Люди не жуют, жуют только коровы {`bridging`, тип `CONTRAST` на узел «человек»}.

- отношение объекта и его функции/позиции (значения `FUNCT_P` и `P_ FUNCT` атрибута `informal-type` в зависимости от направления отношения). Напр. *школа — учитель, министр — министерство* и др.
- остальное (значение `REST` атрибута `informal-type`). В эту группу включаются отношения, которые не были описаны выше, но которые, возможно, будут позже уточнены и выделены в новые группы. Предполагается, что лингвисты-аннотаторы не будут загромождать этот тип парами, которые просто как бы то ни было семантически связаны, а помещать туда только потенциально классифицируемые случаи. В частности к ним относятся отношения место — житель (*Москва — москвич*), автор — творение, вещь — хозяин, родственные отношения (*дед — внук*), некоторые предикатно-аргументные отношения (*предпринимательство — предприниматель, спор — участник конфликта* и др.) а также некоторые релевантные для связности текста равнолексемные некореферентные пары (*случайность — еще одна случайность*)

* * *

Разметка именной текстовой кореференции и ассоциативной анафоры проводится в настоящее время автором данного доклада и тремя аннотаторами с лингвистическим образованием и знаниями в области теории референции и дискурса. Разметка проводится с помощью программы для аннотирования корпусных данных TrEd (од *tree editor*), разработанная на ÚFAL MFF UK, с использованием специально созданных приложений для разметки кореференции. Разметка проводится в основном вручную непосредственно на дереве зависимостей или на тексте (по желанию разметчика). Кроме того, было разработано несколько программ, упрощающих и ускоряющих процесс аннотирования: предварительное выделение лемм, совпадающих с актуальной, указание кореферентных связей данного узла и др. К концу 2008 года было размечено 7000 предложений.

Литература

1. *Hajičová E., Hajič J., Hlaváčová J., Klimeš V., Mírovský J., Pajas P., Štěpánek J., Vidová-Hladká B., Žabokrtský Z.* PDT 2.0 — Guide. UFAL & CKL, 2006 Доступно на http://ufal.mff.cuni.cz/pdt2.0/
2. *Hirschman L.* MUC-7 coreference task definition version 3.0. // Proc. of the 7th Message Understanding Conference под ред. Chinchor N. 1998. Доступно на http://acl.ldc.upenn.edu/muc7/co_task.html
3. *Kučová L.* и др. Anotování koreference v Pražském závislostním korpusu. ÚFAL/CKL Technical Report TR-2003-19. 2003
4. *Mikulova M.* и кол. Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. — Institute of formal and applied linguistics, Charles University, Prague, 2005.
5. *Nědolužko A.* Zpráva k anotování rozšířené textové koreference a bridging vztahů v Pražském závislostním korpusu. (Report about the annotation of the extended text-coreference and bridging relations in Prague Dependency Treebank.). Technical report. Institute of formal and applied linguistics, Charles University, Prague. 2007 Доступно на http://ufal.mff.cuni.cz/~nedoluzko/koref_anot/manual_RK_kratky.pdf
6. *Passonneau R. J.* Instructions for Applying Discourse Reference Annotation for Multiple Applications (DRAMA), 1997
7. *Poesio M.* The MATE/GNOME Scheme for Anaphoric Annotation, Revisited // Proc. of SIGDIAL, Доступно на Boston, April. 2004. Доступно на http://cswww.essex.ac.uk/staff/poesio/publications/SIGDIAL04.pdf
8. *Недолужко А., Гаич Я.* Синтаксически аннотированный корпус чешского языка. // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог». Выпуск 7 (14) 2008 С.400-406 Падучева Е. В. О референции языковых выражений с непредметным значением. // НТИ, сер. 2, N 1, 1986.
9. *Падучева Е. В.* Высказывание и его соотнесенность с действительностью. М.: Наука, 1985.
10. *Степанов Ю. С.* Имена, предикаты, предложения (семиологическая грамматика). М.: Едиториал УРСС, 2004