СИНТАКСИЧЕСКИ АННОТИРОВАННЫЙ КОРПУС ЧЕШСКОГО ЯЗЫКА
THE PRAGUE DEPENDENCY TREEBANK

Недолужко А. *(nedoluzko@ufal.mff.cuni.cz),* ***Гаич Я.*** *(*hajic@ufal.mff.cuni.cz*),* ***и кол.***

Институт формальной и прикладной лингвистики, физико-математический факультет, Карлов университет, Прага, Чехия (ÚFAL MFF UK)

The Prague Dependency Treebank (PDT 2.0) – это корпус текстов чешского языка, аннотированный на трех связанных между собой уровнях – морфологическом (2 млн словоупотреблений), поверхностно-синтаксическом (1.5 млн) и глубинно-синтаксическом (0.8 млн). На глубинно-синтаксическом уровне аннотируется также актуальное членение предложений и именная кореференция. PDT 2.0 основан на пражской лингвистической традиции, адаптированной к требованиям современной компьютерной лингвистики. Аннотация корпуса проводится частично автоматически.

Помимо обширного корпуса чешских текстов разрабатывается проект параллельных текстов на чешском и английском языках (The Prague Czech-English Dependency Treebank), где подобным образом аннотируются тексты из Wall Street Journal и их переводы на чешский язык. Целью проекта является подготовка текстовой базы для обучения компьютера машинному переводу.

В реферате я представлю общую схему аннотации с особым акцентом на глубинно-синтаксический уровень, расскажу о системе синтаксических функторов узлов на этом уровне и словаре моделей управления предикатов, встроенном в проект, а также отвечу на все возникшие вопросы.

THE PRAGUE DEPENDENCY TREEBANK

Nedoluzhko A. *(nedoluzko@ufal.mff.cuni.cz),* ***Hajič J.*** *(*hajic@ufal.mff.cuni.cz*)& Co.*

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

The Prague Dependency Treebank 2.0 (PDT 2.0) contains a large amount of Czech texts with complex and interlinked morphological (2 million words), syntactic (1.5 MW) and complex semantic annotation (0.8 MW); in addition, certain properties of sentence information structure and coreference relations are annotated at the semantic level. PDT 2.0 is based on the long-standing Praguian linguistic tradition, adapted for the current Computational Linguistics research needs. The corpus itself uses the latest annotation technology.

Besides the large corpus of Czech, a corpus of Czech-English parallel resources (The Prague Czech-English Dependency Treebank) is being developed. English sentences from the Wall Street Journal and their translations into Czech are being annotated in the same way as in PDT 2.0. This corpus is suitable for experiments in machine translation, with a special emphasis on dependency-based (structural) translation.

In the report, the basic annotation scheme is represented, with special reference to complex semantic (tectogrammatical) level. The system of syntactic functors and valency lexicon VALLEX are also discussed.

1. Общие сведения

Синтаксически аннотированный корпус чешского языка (PDT) – это проект лингвистического (морфологического, синтаксического, семантического, прагматического и др.) аннотирования текстов, разрабатываемый в настоящее время в Институте формальной и прикладной лингвистики физико-математического факультета Карлова университета в Праге. Последняя версия проекта, PDT 2.0, содержит большое количество чешских текстов (2 млн. словоупотреблений) с аннотацией (взаимосвязанной) на трех уровнях – морфологическом (2 млн. слов), поверхностно-синтаксическом (1.5 млн. слов) и глубинно-синтаксическом (0.8 млн. слов). Корпус использует самые современные способы аннотации (раздельная аннотация уровней с использованием XML, RelaxNG). К корпусу также прилагается отдельная поисковая программа Netgraph, позволяющая производить сложный поиск по многим параметрам и собирать материал и статистические данные для лингвистических исследований.

Аннотирование синтаксических уровней производится вручную на основе предварительных автоматических аннотаций, т.е. фактически аннотирующий лингвист просматривает уже готовую аннотацию, дополняет ее и исправляет ошибки. Аннотирование синтаксических уровней проводится с помощью специальной программы для аннотирования корпусных данных TrEd (од *tree editor*), разработанная на ÚFAL MFF UK. Аннотирование вручную проводится аннотаторами с лингвистическим образованием, причем регулярно проводится тест на т.наз. «соответствие аннотаторов», т.е. все аннотаторы, работающие на данном проекте, аннотируют одни и те же тексты, на которых затем проводится автоматическая проверка соответствия.

Лингвистическая основа PDT восходит к традициям пражской лингвистической школы и функционально-грамматическому описанию языка, разработанному в шестидесятых годах двадцатого века чешским лингвистом П. Сгаллом и его учениками.

PDT - один из нескольких десятков проектов синтаксически аннотированных корпусов, разрыбатываемых в настоящее время в мире. Идейным вдохновителем проекта послужил американский PennTreebank (http://www.cis.upenn.edu/~treebank), однако со структурной точки зрения он значительно отличается от PDT и разработан на основе принципа непосредственных составляющих. Лингвистически близким PDT является разработка И.Богуславского и система уровней ЭТАПа-3, но в PDT несравнимо большую роль играет статистика, иначе работает система синтаксических отношений, больше объем обработанного автоматически и вручную материала и т.д. С т.з. количества синтаксически обработанного материала PDT можно сравнить с корпусом датских текстов Danish Dependency Treebank – 5500 синтаксически аннотированных деревьев (http://www.id.cbs.dk/~mbk/treebank), португальских текстов - The Floresta Sintá(c)tica project, 10000 деревьев (http://acdc.linguateca.pt/treebank/info_floresta_English.html), турецких текстов – METU-Sabanci Turkish Treebank (http://www.ii.metu.edu.tr/~corpus/treebank.html) и др. Несомненным преимуществом PDT является комбинация большого количества аннотированных текстов с богатой лингвистической информацией, в т.ч. выходящей за рамки одного предложения (аннотация кореференции, актуального членения, сочинительных конструкций и др.)

2. Уровни аннотации

Аннотирование проводится на трех уровнях – морфологическом, поверхностно-синтаксическом и глубинно-синтаксическом. В действительности существует еще нулевой уровень основного текста, где всем элементам (слова, числа, знаки препинания) присваиваются идентификаторы. На рис. 1 изображена связь между уровнями: так, как они аннотируются в PDT 2.0. Это разбор чешского предложения *Byl by šel dolesa (Шел бы в лес)*, содержащее глагол «идти» в сослагательном наклонении в прошедшем времени (*Byl by šel*) и опечатку (*dolesa «в лес»* написано слитно, должно быть *do lesa*).

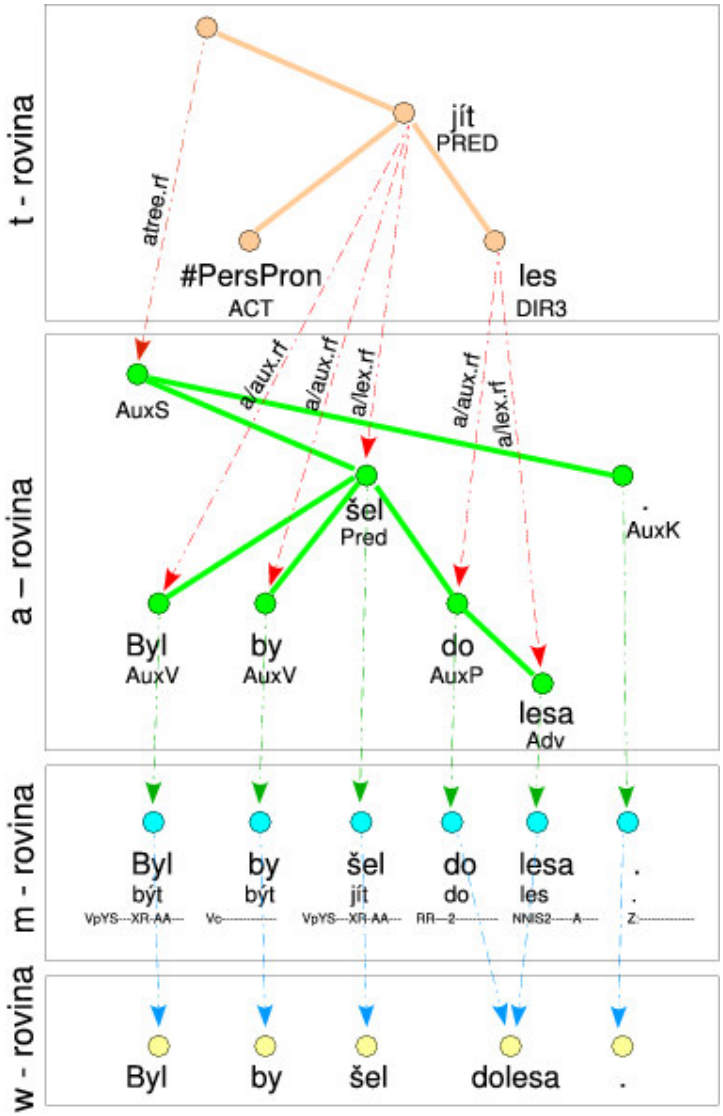


Рис. 1 Связи уровней аннотации в PDT 2.0

2.1 Морфологический уровень.

Здесь словоупотреблениям нулевого уровня присваивается некоторое количество атрибутов, из которых самыми важными являются морфологические: lemma и tag. Атрибут lemma представляет собой имя лексемы данного слова и однозначно соотносит его с морфологическим словарем. Атрибут tag содержит 15 позиций морфологической информации (часть речи и все актуальные для нее морфологические характеристики, напр. NNIS2-----A----). Пример аннотации на морфологическом уровне рассмотрен ниже.

Аннотация морфологического уровня проводилась группой из семи аннотаторов, и была разделена на два этапа. На первом этапе каждый текст был предварительно аннотирован морфологическим анализатором. Затем два аннотатора, независимо друг от друга, проконтролировали правильность атрибутов lemma и tag. На втором этапе все несоответствия между этими двумя аннотаторами были разрешены третьим, контролирующим аннотатором. После окончания аннотирования поверхностно-синтаксического уровня была проведена еще одна ревизия, для проверки соответствия предлогов и падежей существительных, именного согласования и т.д.

2.2 Поверхностно-синтаксический уровень (ПСУ)

Здесь структура предложения представлена в виде ориентированного дерева с помеченными связями (ребрами) и узлами. Каждому элементу морфологического уровня соответствует узел поверхностно-синтаксического дерева, отношения между элементами выражены связывающими их ребрами. Тип отношения определяется типом ребра – большинство ребер отражают отношение зависимости, но есть и другие отношения, напр. координация, аппозиция, знаки препинания и др.

Каждому узлу приписывается шесть атрибутов. Атрибут `id` содержит однозначный в рамках PDT 2.0 идентификатор узла, который связывает его с глубинно-синтаксическим уровнем. Линейный порядок узлов отражается в атрибуте `ord`. Функция ребра по техническим причинам отображается в атрибуте `afun` у нижнего узла. Атрибуты `is_member` и `is_parenthesis_root` используются для указания на сочинительные конструкции и выражения в скобках. Атрибут `m.rf` связывает узел с соответствующим элементом на морфологическом уровне. Пример аннотации на ПСУ рассмотрен ниже.

Все данные ПСУ аннотировались группой из шести аннотаторов – сначала вручную, затем на основе предварительной автоматической аннотации. По окончании аннотирования проводились всевозможные контрольные тесты, найденные ошибки были вручную проверены и исправлены.

2.3 Глубинно-синтаксический (тектограмматический) уровень (ГСУ)

Структура ГСУ – дерево, где каждому узлу, кроме технического корня, присвоено 39 атрибутов. В зависимости от типа узла (атрибут `nodetype`) заполняется определенное подмножество этих атрибутов. Наибольший интерес представляют следующие атрибуты:

Атрибут `functor` – описывает тип ребра, ведущего от узла к его предку – зависимость или другое техническое отношение. Значениями этого атрибута могут быть функторы для актантов (ACT – агенс, PAT – пациенс, ADDR – адресат и др.), функторы корней независимых клауз (PRED – главный предикат предложения, DENOM – именной корень клаузы, PAR – корень выражения в скобках), функторы для корней сочинительных конструкций (CONJ – сочинительная конструкция, ADVS – противительная конструкция и др.), функторы места (LOC - где, DIR1 - откуда, DIR2 – каким путем, DIR3 - куда) и времени (TWHEN - когда, TTILL – до какого времени, TSIN – с какого времени, TPAR – в течение какого времени и др.) и другие. Всего на данный момент для аннотирования чешского языка используется 67 функторов, распределенных на 12 групп.

Атрибут `t_lemma` содержит имя лексемы на глубинно-синтаксическом уровне.

16 атрибутов используется для описания грамматических свойств узла. Эти атрибуты обозначены префиксом `gram` (напр., атрибут `gram/sempos` – семантическая часть речи, имеющий далее 19 значений: n.denot - семантическое существительное, adj.denot - семантическое прилагательное, v - глагол и т.д.; атрибут `gram/verbmod` содержит информацию о модальности предложения и т.п.)

Так как тектограмматическая структура, также как и ПСУ, основана на синтаксических зависимостях, для конвертирования поверхностно-синтаксических деревьев в предварительные глубинно-синтаксические были использованы автоматические методы. Все полученные таким образом деревья были затем вручную обработаны аннотаторами, которые дополнили большое количество недостающей информации и исправили ошибки.

2.3.1 Словарь моделей управления VALLEX. На ГСУ предикатам присваивается модель управления из связанного с TrEd-ом словаря валентностей **VALLEX**. Это электронный словарь, содержащий примерно 2730 лексем. Словарная статья включает как минимум одну модель управления с указанием обязательных актантов и их возможных синтаксических реализаций, а также с примерами их употребления. Например, представление глагола rozumět (понимать) в pdf-версии словаря выглядит так:

rozumět *impf* v

1 (*vyznat se; chápat*) ACT(1) PAT(3|zda|že|cont) ◊*rozumí úloze; nerozuměl, zda to má nebo nemá udělat; rozumíš už, co se stalo?; rozumí dobře anglicky; matka dceři rozumí* ✠ rcp: ACT-PAT; class: mental action

2 (*rozlišovat; znát*) ACT(1) ADDR(3) PAT(4|zda|že|cont) ¿MANN ◊*rozumí mu každé slovo; Rozumím vám správně, že rozpočet v parlamentě podpoříte? (ČNK)* ✠ rcp: ACT-ADDR; class: communication

3 idiom (*chápat*) ACT(1) PAT(4|že) EFF(7|pod+7) ◊ *rozumí tím / pod tím příměří* ✠ rfl: pass; class: mental action

При аннотировании ГСУ модель управления должна быть присвоена глаголам и отглагольным прилагательным на *-ní* (типа *koupaní - купание*) и *–tí* (*mýtí - мытье*). Предикативам – представителям других частей речи модель управления пока последовательно не присваивается.

Помимо грамматической структуры зависимостей, на тектограмматическом уровне имеется также информация об актуальном членении предложений и о кореференции, которая аннотировалась отдельно.

2.3.2 Актуальное членение. Аннотирация актуального членения основана на двух традиционных чешских концепциях: В. Матезиуса о теме-реме и контекстной связанности и Я. Фирбаса о функциональной перспективе предложения. В аннотации PDT 2.0 фиксируется контекстная связанность (данность, известность) узлов и функциональная перспектива предложения. Информация о тематических и рематических блоках должна автоматически высчитываться на основе этих данных. Контекстная связанность представлена значениями атрибута `tfa` (topic-focus articulation) – *t* (данное), *f* (новое) и *c* (контраст) и аннотируется вручную, отдельно для каждого узла. Атрибут `deepord` используется для обозначения глубинного порядка узлов, основанного на функциональной перспективе предложения. Таким образом, в глубинно-синтаксическом представлении порядок узлов слева направо обозначает степень их функциональной динамичности – от наименее к наиболее динамичному элементу.

2.3.3 Кореференция. В аннотации PDT 2.0 кореференция делится на грамматическую и текстовую. Другие случаи кореференции, такие как экзофорическая отсылка и отсылка к двум и более предложениям, рассматриваются отдельно. В качестве антецедента может выступать терминальный узел дерева, поддерево (отсылка на корень поддерева) или несколько поддеревьев (отсылка на корневые узлы этих поддеревьев)

В случае грамматической кореференции антецедент высчитывается на основании грамматических правил языка, к ней относится кореференция возвратных местоимений (в чешском языке возвратное местоимение – всегда отдельная клитика), относительных местоимений (напр. *человек, который* пьет; *в городе, где* мне так понравилось и др.) и др. Грамматическая кореференция практически никогда не переходит границ предложения, ее всегда можно представить как отсылку одного узла к другому, следовательно ее аннотирование легко автоматизируется.

Текстовая кореференция аннотируется в PDT 2.0 только в том случае, если в качестве анафорического члена выступают личные и посессивные местоимения третьего лица, указательное местоимение этот в субстантивной функции и актуальный эллипсис этих местоимений, восстанавливаемый на ГСУ. Текстовая кореференция может легко переходить границы предложения, и ее определение часто зависит от знания контекста, поэтому ее аннотирование проводилось вручную.

Для аннотирования кореференции используется id антецедента, к которому отсылает id узла анафоры. Атрибуты `coref_text.rf`, и `coref_gram.rf` содержат id кореферентных узлов соответствующих типов. Атрибут `coref_special` несет информацию об особых случаях кореференции.

В настоящее время разрабатывается проект расширенного аннотирования кореференции, где текстовая кореференция будет дополнена случаями, когда в качестве анафорического повтора выступают другие части речи (прежде всего существительные – повторение данной ИГ, синонимы, гиперонимы и т.д.), но при этом сохраняется тождество референтов. Кроме того, планируется включить в аннотацию случаи т.наз. bridging anaphora, где референты антецедента и анафорического «повтора» уже не тождественны, но семантически связаны. Над этой темой сейчас работает автор данного доклада.

3. Пример предложения, аннотированного на трех уровнях аннотации в PDT 2.0

Některé kontury problému se však po oživením Havlovým projevem zdají být jasnější. – Некоторые контуры проблемы однако после оживлением выступления Гавела кажутся понятнее

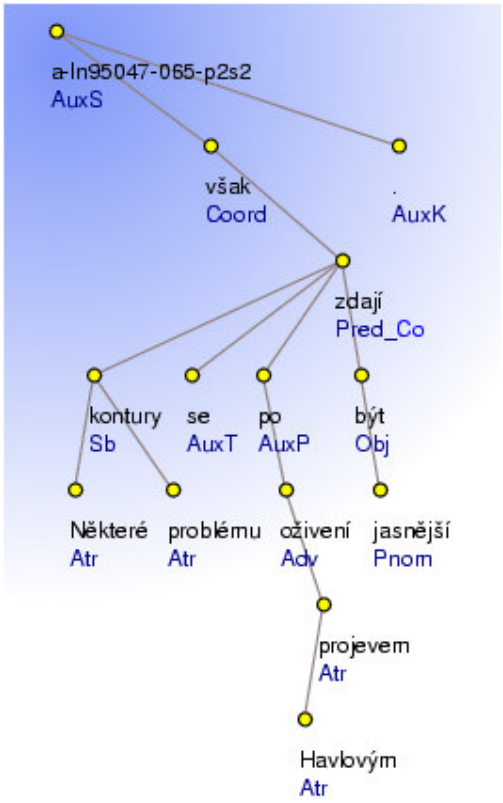
3.1. Нулевой уровень слов:

Některé	*kontury*	*problému*	*se*	*však*	*po*	*oživením*	*Havlovým*	*projevem*	*zdají*	*být*	*jasnější*	*.*
некоторый adj. masc Npl	контур noun, masc, Npl	проблема masc, Gsg	возвр. «ся» pron.	однако adv	по prep	оживление noun,neutr, DSg	Гавлов adj-poss, masc, ISg	выступления noun, neutr ISg	кажут(ся) verb, ind,act, praes. 3Sg	быть inf	ясный ср.степ.	.

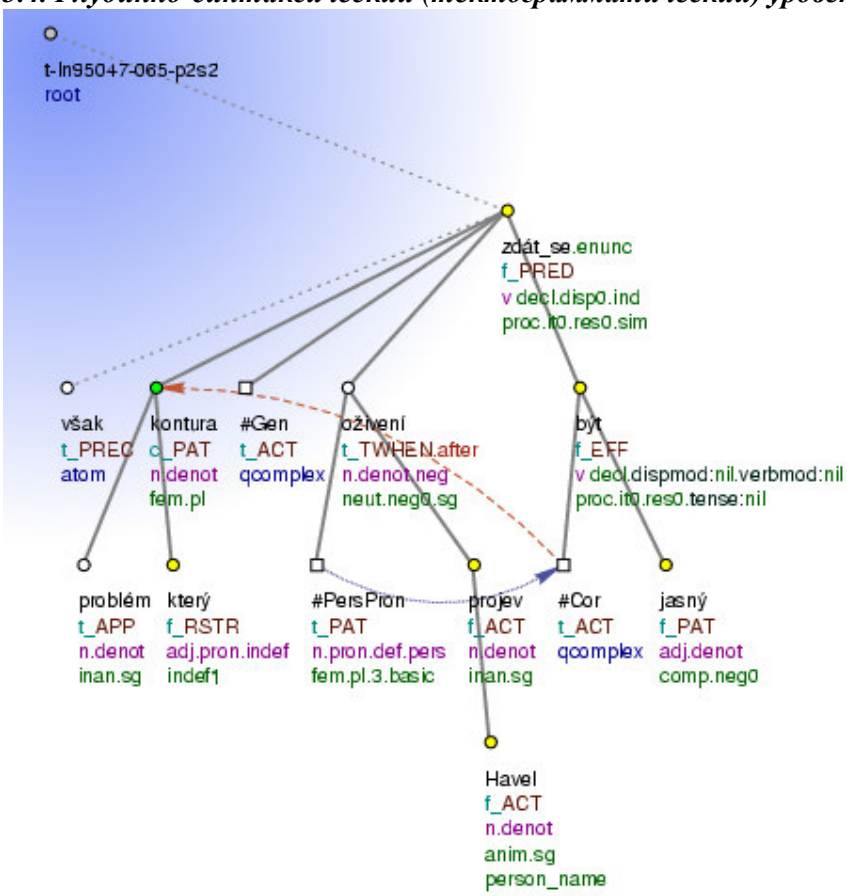
3.2. Морфологический уровень

словоформа	лемма	морфологический тег
Některé	*některý*	PZFP1----------
kontury	*kontura*	NNFP1-----A----
problému	*problém*	NNIS2-----A----
se	*se*_^(zvr._zájmeno/částice)	P7-X4----------
však	*však*	J^----------
po	*po*-1	RR--6----------
oživení	*oživení*_^(*3it)	NNNS6-----A----
Havlovým	*Havlův*_;S_^(*3el)	AUIS7M---------
projevem	*projev*	NNIS7-----A----
zdají	*zdát*	VB-P---3P-AA---
být	*být*	Vf--------A----
jasnější	*jasný*	AAFP1----2A----
.	*.*	Z:----------

3.3. Поверхностно-синтаксический уровень.



3.4. Глубинно-синтаксический (тектограмматический) уровень



***

В недавнем прошлом проект PDT был дополнен Пражским арабским синтаксически аннотированным корпусом (Prague Arabic Dependency Treebank, http://www.ldc.upenn.edu) и параллельным чешско-английским корпусом (Prague Czech-English Dependency Treebank, http://ufal.mff.cuni.cz/pcedt). Арабский проект подтверждает, что разработанная на чешском языке система может работать и на типологически несходном языке. Синтаксически аннотированный параллельный чешско-английский корпус разрабатывается на основе аннотрирования вручную текстов из журнала Wall Street Journal, которые ранее использовались для корпуса Penn Treebank 3. В настоящее время параллельно аннотируется примерно 21600 предложений на английском языке и их переводы на чешский. Целью проекта является подготовка текстовой базы для обучения компьютера машинному переводу с чешского на английский и обратно.

Проект PDT имеет и более далеко идущие планы. Рассматриваются несколько вариантов: пополнение PDT разговорными текстами, детализация имеющейся аннотации (в основном в области аннотации кореференции, информационной структуры и дискурса), аннотация других типологически отличающихся языков, аннотация вручную глубинно-синтаксического уровня на параллельных чешских и английских текстах, разработка новых уровней аннотации и т.д. По большинству из этих проектов уже ведутся разработки.

Литература:

1. Čmejrek M., Cuřín J., Havelka J., Hajič J., Kuboň V. Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation, In 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal. Доступно на http://ufal.mff.cuni.cz/pcedt/doc/papers/lrec2004_pcedt.pdf. 2004.
2. Hajič J., Hajičová E., 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/
3. Mikulova M. и кол. Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka Institute of formal and applied linguistics, Charles University, Prague, 2006.
4. 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
5. Žabokrtský, Z.; Lopatková, M.: Valency Frames of Czech Verbs in VALLEX 1.0. // In Frontiers in Corpus Annotation. Proceedings of the Workshop of the HLT/NAACL Conference, pp. 70--77. 2004