Разметка анафорических отношений и дискурса в синтаксически аннотированном корпусе чешского языка (Prague Dependency Treebank)

Анна Недолужко

Москва, май 2015

План лекции

* Более или менее техническая лекция – представим текстовую разметку в PDT 3.0
* Общие сведения о корпусах и об уровнях разметки
* Глубинно-синтаксический уровень разметки
* Разметка на уровне текста:
  * кореферентность,
  * ассоциативная анафора,
* задание
* TrEd и выгоды/невыгоды разметки на деревьях
* Разметка на уровне текста:
  * дискурсивная разметка

PDT 3.0
Синтаксически аннотированный корпус чешского языка

PCEDT 2.0+
Синтаксически аннотированный параллельный чешско-английский корпус

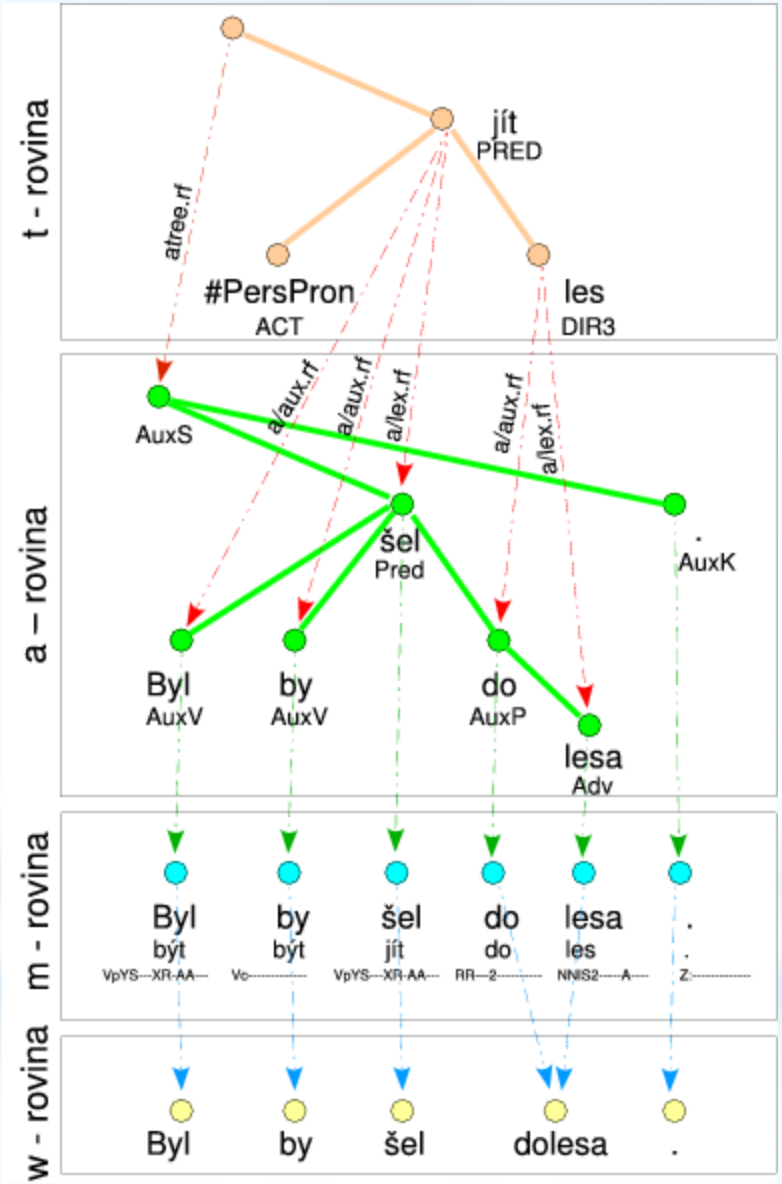
PDT 3.0 (Синтаксически аннотированный корпус чешского языка)

* The Prague Dependency Treebank [Bejček et al. 2013]
* чешские газетные тексты,
* ca. 50000 предложений(3165 документов, 833195 токенов)
* размеченные на морфологическом (m-layer), поверхностно-синтаксическом (shallow syntactic, a-layer) и глубинно-синтаксическом (тектограмматическом, t-layer) уровнях
* t-layer включает:
  * гл.-синтаксическую разметку слов и конструкций
  * argument structure description based on a valency lexicon
  * восстановление синтаксического эллипсиса
  * разметка кореферентности (местоимения, нули, NP вкл. различия на КР/ГЕН)
  * разметку ассоциативной анафоры
  * дискурсивную разметку (дискурсивные маркеры (коннекторы), аргументы связанные этими коннекторами и типы отношений между ними)

PCEDT 2.0 (Синтаксически аннотированный параллельный чешско-английский корпус)

- Prague Czech-English Dependency Treebank [Hajic et al., 2012]
- тексты English Wall Street journal переведенные на чешский по предложениям
- 1.2 млн слов, ок. 50,000 предложений для каждого языка
- размеченные на морфологическом (m-layer), поверхностно-синтаксическом (shallow syntactic, a-layer) и глубинно-синтаксическом (тектограмматическом, t-layer) уровнях
- sentence-aligned, word-aligned
- t-layer включает:
  - гл.-синтаксическую разметку слов и конструкций
  - восстановленную аргументную структуру предикатов на основе словаря валентностей
  - разметку кореферентности
  - восстановленные нули

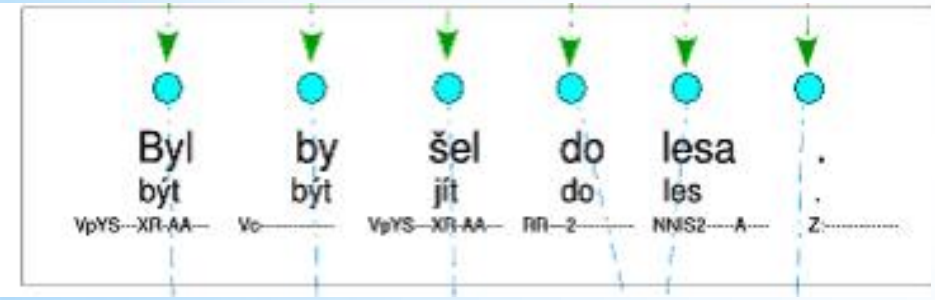
Byl by šel dolesa.



(Он) пошёл бы влес.

*Морфологический уровень

(Он) пошёл бы в лес.

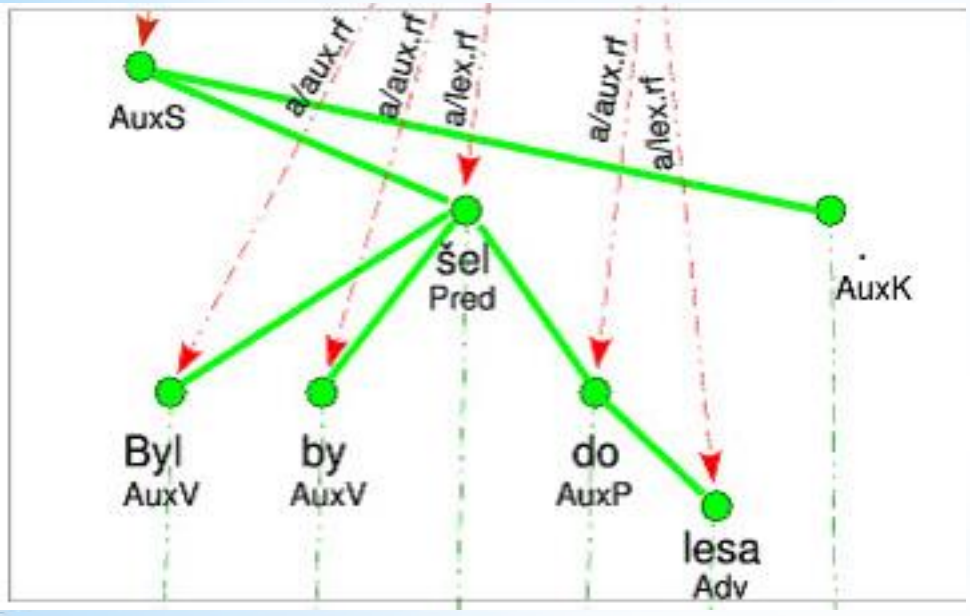


атрибуты:

* атрибут lemma
* атрибут tag - 15 позиций,

  напр. **NNIS2-----A----**
* и др.

*Поверхностно-синтаксический уровень

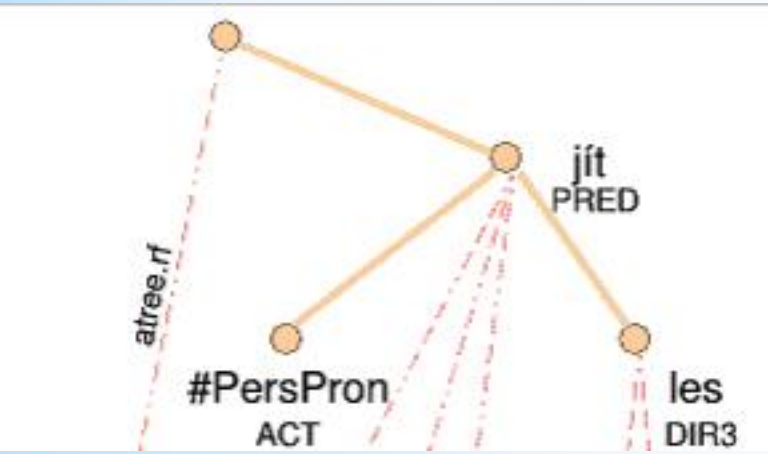


атрибуты (6):

- *id
- *ord
- *afun
- *is_member
- *is_parenthesis_root
- *m.rf

Глубинно-синтаксический уровень

(Он) пошёл бы влес.



атрибуты (39)

- functor (ACT, PAT, ADDR, PRED, DENOM, PAR, CONJ, LOC, DIR1, DIR2, TWHEN, TTILL и др.)
- t_lemma
- gram/sempos, gram/verbmod

*Словарь моделей управления VALLEX

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

cz. rozumět = ru. понимать

Разметка на уровне текста

Глубинно-синтаксический уровень PDT 3.0 включает разметку лингвистических явлений с т.з. структуры дискурса и когеренции.

- разметка **кореферентности,**
- разметка ассоциативной анафоры (бриджинг),
- восстановление **синтаксического эллипсиса**
- разметка **актуального членения**(tfa)
- **дискурсивная** разметка: дискурсивные маркеры (коннекторы), аргументы связанные этими коннекторами и типы отношений между ними

Разметка на уровне текста

Глубинно-синтаксический уровень PDT 3.0 включает разметку лингвистических явлений с т.з. структуры дискурса и когеренции.

- разметка **кореферентности,**
- разметка ассоциативной анафоры (бриджингов),
- восстановление **синтаксического эллипсиса**
- разметка **актуального членения**(tfa)
- **дискурсивная** разметка: дискурсивные маркеры (коннекторы), аргументы связанные этими коннекторами и типы отношений между ними

Кореферентность

*Кореферентность

* **грамматическая (структурная) кореферентность** –
  * правила языка → антецедент,
  * в рамках одного предложения
* **текстовая кореферентность** –
  * не ограничена грамматикой, влияет контекст, знания о мире и т.д.
  * реализутеся прономинализацией, согласованием, повторами, синонимами, переформулированием, гипоним/гиперонимы и т.д.
  * часто пересекает границы предложения

Грамматическая (структурная) кореферентность

* Аргумент зависимого глагола в контролирующих конструкциях
  – *Ваня хочет* PRO *поцеловать Машу*.
* Кореферентность возвратных местоимений
  – *Ваня уважал себя.* *Ваня побрился*.
* Кореферентность относительных средств
  – *Ваня, который опоздал, сильно извинялся*.
* Кореферентность с глагольными дополнениями с двумя зависимостями
  – *John saw Mary* PRO *stand on the windowsill and cry.*
  (– Ваня увидел Машу, стоящую на подоконнике и плачущую.)
* Кореферентность в возвратных конструкциях
  – *Ваня и Маша поцеловались* PRO .

Текстовая кореферентность

- личные и поссессивные местоимения (*Ваня оставил Машу. Его мама просила, чтобы вечером он был дома*),
- указательные местоимения в субстантивной функции (*Это значит, что на самом деле он не очень любил Машу...*)
- Эллипсис (pro) (*… и гораздо больше pro уважал свою мать.*)
- существительные (*Ваня просил маму объяснить ему, как ему вести себя с Машей, но мама сочла эту просьбу сына абсурдной.*)
- наречия места и времени (*Тогда Ваня попросил маму съездить за Машей в Новгород, но она решила туда не ездить.*)
- некоторые прилагательные (от ИС=NE) (*В конце концов, Маша сама приехала в Москву и московская жизнь ей очень понравилась.*)
- отсылка к клаузам, действиям (*Маша предложила Ване сходить вместе в театр, но Ваня ее предложение проигнорировал.*)
- отсылка к сегментам текста, большим, чем одно предложение – особый тип отношения, не маркирующий антецедент (segm) (*На следующий день Маша захотела посетить его мать. Потом она предложила сходить поплавать. Ее последним желанием было посмотреть на центр города. Ваня от всего этого отказался.*)

*Текстовая кореферентность - типы

* **SPEC(ific)** кореферентные отношения конкретно-референтных ИГ

  (*Ваня$_a$ попросил маму$_b$ объяснить$_c$ ему$_a$, как ему$_a$ вести себя с Машей, но мама$_b$ сочла эту просьбу$_c$ своего$_b$ сына$_a$ абсурдной*)

* **GEN(eric)** кореферентные отношения родовых и других нереферентных ИГ.

  (Маша предложила Ване сходить в зоопарк посмотреть на зверей. Она думала, что посмотрев на зверей, Ваня поймет, как зверски он вел себя по отношению к ней.)

Ассоциативная анафора (бриджинг)

Ассоциативная анафора (бриджинг)

* некореференткое семантическое отношение (*дом - крыша*)
* имеет значение для связности текста
* не сохраняется цепочка
* размечаем примерно в смысле Clark (1975), но только некоторые типы

Ассоциативная анафора (бриджинг)

- **часть – целое** (PART_WHOLE и WHOLE_PART)

  Россия – Владимирская область – Юрьев Польский

- **множество — подмножество/элемент множества** (SET_SUB и SUB_SET)

  студенты – несколько студентов – студент

- **объект — уникальная функция на нем** (P_FUNCT и FUNCT_P)

  премьер-министр - правительство, футбольная команда - тренер

- **отношение семантического и прагм. контраста** (CONTRAST)

  Люди не жуют, жуют только коровы.

- эксплицитная **анафора без кореферентности** (ANAF)

  «... а над небом красовалась огромная радуга...» Валера приложил палец к этому слову, чтобы не забыть, где он остановился, и посмотрел на Игоря. ...

  У Вас замечательные студенты! Таких бы студентов и к нам на кафедру...

- **другое** (REST) – 5 маленьких групп

  члены семьи *(дед - сын)*, место - житель, автор - творение, действие – партиципант (*проституция - проститутка*)

Экзофорическая отсылка

Экзофорическая отсылка – за рамки текста, к ситуации – special attribute coref_special, type exoph

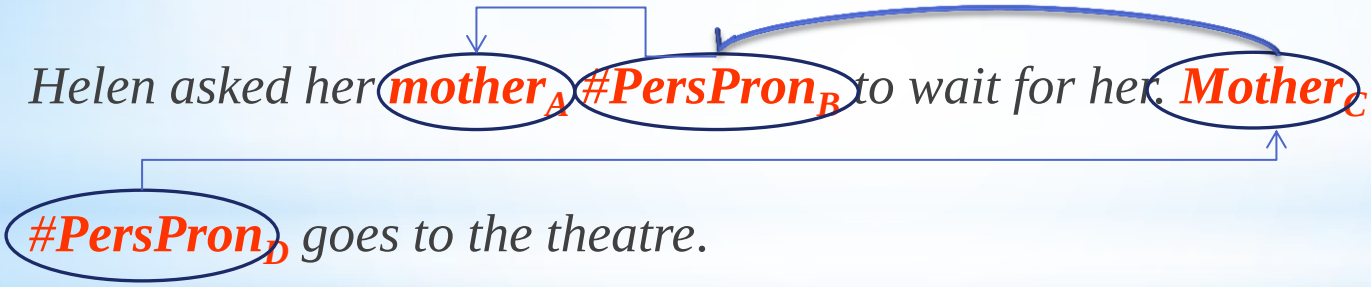
* *Dokončeny by měly být* [...] *na sídlišti Barrandov v těchto dnech.(= It should be finished* [...] *in Barrandov district in these days [meaning, in the recent days])*
* *A tu se dostáváme zpět k počátku tohoto textu . (= With this, we come back to the beginning of this text)*
* *Informace v tomto přehledu jsou bezplatnou službou podnikatelům. (=The information in this report is a free service to businessmen.)*

*Принципы разметки кореферентности -1

- кореф. цепочки – отсылка к ближайшему антецеденту
- принцип максимальной длины цепочек

Пример:

Helena poprosila svou ***maminku***$_A$*, aby* ***#PersPron***$_B$ *na ni počkala.*
Matka$_C$ *řekla, že* ***#PersPron***$_D$ *jde do divadla.*



получается цепочка: A <= B <= C <= D

*Принципы разметки кореферентности -2

- максимальные “scope” маркабулы: все поддерево
- “сотрудничество” с глубинно-синтаксическим уровнем: не размечаем кореферентность между членами аппозиций и предикативных конструкций
- при прочих равных смотрим на когерентность текста
- кореференция важней, чем bridging: при необходимости выбора выбираем кореферентность

  Mary – John – children in the class – Mary and John

- в первую очередь размечаем кореферентность, а не анафорическое отношение

* (1) Argentina said it will ask creditor banks to halve its foreign debt of $64 billion -- the third-highest in the developing world.
* (2) The declaration by Economy Minister Nestor Rapanelli is believed to be the first time such an action has been called for by an Argentine official of such stature.
* (3) The Latin American nation has paid very little on its debt since early last year.
* (4) “Argentina aspires to reach a reduction of 50% in the value of its external debt,” Mr. Rapanelli said through his spokesman, Miguel Alurralde.
* (5) Mr. Rapanelli met in August with U.S. Assistant Treasury Secretary David Mulford.
* (6) Argentine negotiator Carlos Carballo was in Washington and New York this week to meet with banks.
* (7) Mr. Rapanelli recently has said the government of President Carlos Menem, who 29 took office July 8, feels a significant reduction of principal and interest is the only way 0 the debt problem may be solved.
* (8) But he has not said before that the country wants half the debt forgiven.

* (1) [Argentina said it will ask creditor banks [to halve [its foreign debt of $64 billion -- the third-highest in the developing world]$_a$]$_b$]$_c$.
* (2) [The declaration by Economy Minister Nestor Rapanelli]$_b$ is believed to be the first time [such an action]$_b$[?] has been called for by an Argentine official of such stature[?].
* (3) The Latin American nation has paid very little on [its debt]$_a$ since early last year.
* (4) "Argentina aspires to #Cor.ACT reach [a reduction of 50% in the value of [its external debt]$_a$]$_c$," Mr. Rapanelli said through [his spokesman]$_d$, [Miguel Alurralde]$_d$.
* (5) Mr. Rapanelli met in August with U.S. Assistant Treasury Secretary David Mulford.
* (6) Argentine negotiator Carlos Carballo was in Washington and New York this week to meet with banks.
* (7) Mr. Rapanelli recently has said the government of President Carlos Menem feels a significant reduction of principal and interest is the only way the [debt]$_a$ problem may be solved.
* (8) But he has not said before that the country wants half [the debt]$_a$ forgiven.

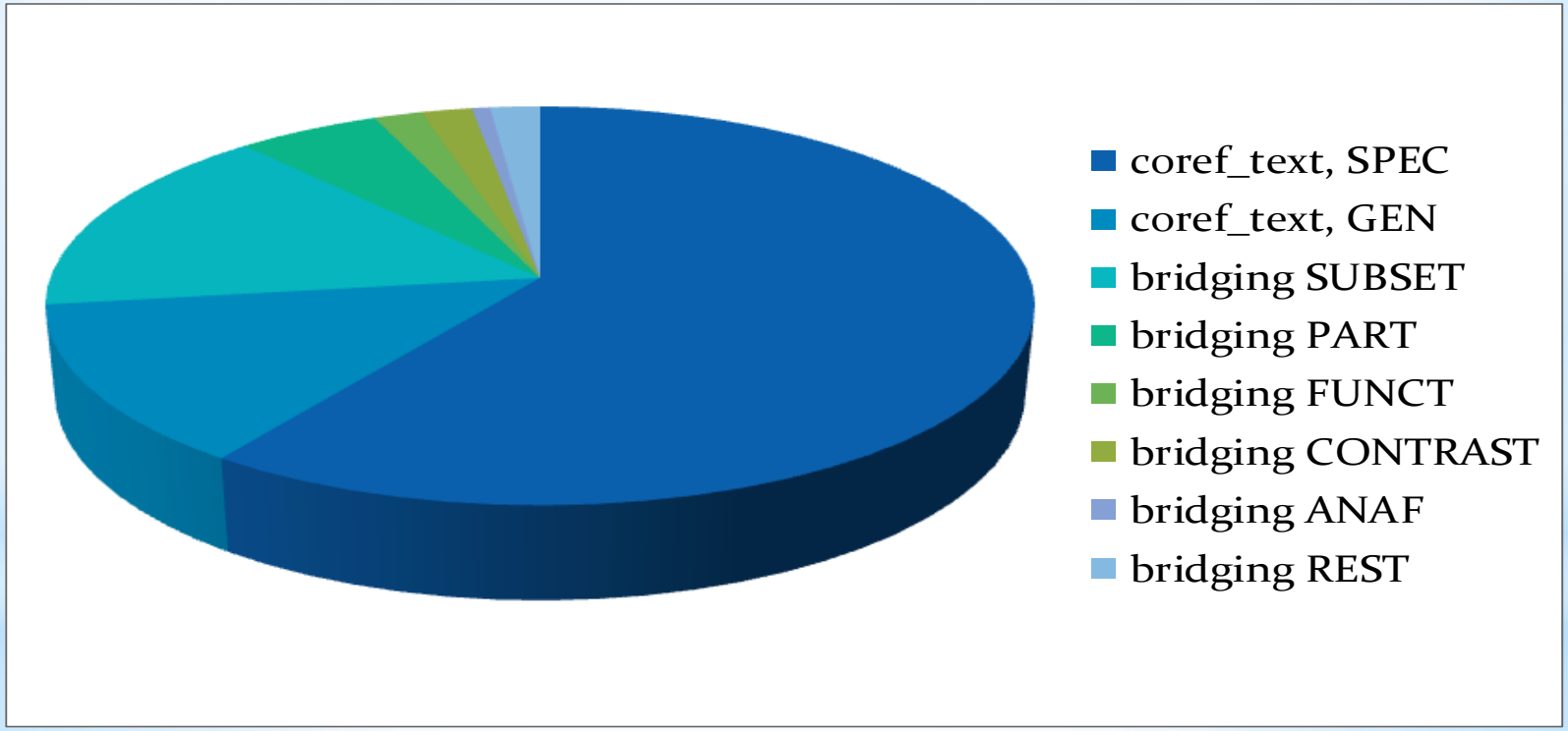
Разметка кореферентности в пражских корпусах

	PDT 3.0	PCEDT	
		английский	чешский
грамматическая кореферентность	ДА	ДА	ДА
кореф. с местоимениями	ДА	ДА	ДА
кореф. с анафорическими нулями	ДА	ДА	ДА
текстовая кореф. – ИГ, конкр.-реф.	ДА	ДА (PCEDT2.0+)	ДА (PCEDT2.0+)
текстовая кореф. – ИГ, ген.	ДА	нет	нет
bridging	ДА	нет	нет

Статистика по разметке кореферентности и бриджинг в PDT

Тип отношения	количество
Всего текстовой кореферентности	86,349
Текстовая кореф. (конкр-реф. NP)	20,243 (мест.)+50,593 (сущ.) = 70,836
Текстовая кореф. (родовые NP)	3,095(мест.)+12,418(сущ.) = 15,513
Всего бриджинг	32,171
бриджинг SUBSET	5,820(SUB_SET) +12,580(SET_SUB) = 18,400
бриджинг PART	1,982(PART_WHOLE)+4,372(WHOLE_PART) = 6,354
бриджинг FUNCT	1,719(P_FUNCT)+418(FUNCT_P) = 2,137
бриджинг CONTRAST	2,238
бриджинг ANAF	802
бриджинг REST	2,212
Процентное соотношение узлов, из которых ведет анафорическая отсылка (кореф. или бриджинг)	17.6%

Статистика по размеченной кореферентности и бриджингу в PDT



Inter-annotator agreement

Проверено документов	39
Проверено предложений	1606 (26,520 tokens)
F-1 на текстовой местоименной кореф. (+ нули)	0,86
F-1 на текстовой именной кореф. с конкр-реф NPs	0,705
F-1 на текстовой именной кореф. с родовыми NPs	0,492
F-1 на бриджинг	0,455
кореф. kappa на соответствие типов	0,759
bridging kappa на соответствие типов	0,889

Обзор систем по разрешениею кореферентности

Тип задачи	корпус	F_1
Grammatical coreference, verbs of control	PDT 2.0	91.5
Grammatical coreference, возвратные местоимения	PDT 2.0	97.1
Grammatical coreference, относительные местоимения	PDT 2.0	99.6
Grammatical coreference, reciprocity	PDT 2.0	94.7
Pronominal coreference, rule-based	PDT 2.0	74.2
Pronominal coreference, perceptron ranking, gold features	PDT 2.0	79.4
Pronominal coreference, perceptron ranking, system features	PDT 2.0	50.3
NP-coreference, specific NPs	PDT 2.0	48.1 (P:59.7, R:40.3)
NP-coreference, generic NPs	PDT 2.0	1.8 (P:20, R:0.9)
bridging relations	PDT 2.0	0
Идентификация анафорического невыраженного субъекта, rule-based	PCEDT 2.0	61.5
Идентификация анафорического невыраженного субъекта, rule-based, с использованием английских текстов	PCEDT 2.0	69.5

Обзор систем по разрешениею кореферентности

Тип задачи	корпус	F_1
Grammatical coreference, verbs of control	PDT 2.0	91.5
Grammatical coreference, возвратные местоимения	PDT 2.0	97.1
Grammatical coreference, относительные местоимения	PDT 2.0	99.6
Grammatical coreference, reciprocity	PDT 2.0	94.7
Pronominal coreference, rule-based	PDT 2.0	74.2
Pronominal coreference, perceptron ranking, gold features	PDT 2.0	79.4
Pronominal coreference, perceptron ranking, system features	PDT 2.0	50.3
NP-coreference, specific NPs	PDT 2.0	48.1 (P:59.7, R:40.3)
(NP-coreference, generic NPs)	PDT 2.0	1.8 (P:20, R:0.9)
(bridging relations) новые features!	PDT 2.0	0
Идентификация анафорического невыраженного субъекта, rule-based	PCEDT 2.0	61.5
Идентификация анафорического невыраженного субъекта, rule-based, с использованием английских текстов	PCEDT 2.0	69.5

Обзор систем по разрешениею кореферентности

Тип задачи	корпус	F_1
Grammatical coreference, verbs of control	PDT 2.0	91.5
Grammatical coreference, возвратные местоимения	PDT 2.0	97.1
Grammatical coreference, относительные местоимения	PDT 2.0	99.6
Grammatical coreference, reciprocity	PDT 2.0	94.7
Pronominal coreference, rule-based	PDT 2.0	74.2
Pronominal coreference, perceptron ranking, gold features	PDT 2.0	79.4
Pronominal coreference, perceptron ranking, system features	PDT 2.0	50.3
NP-coreference, specific NPs	PDT 2.0	48.1 (P:59.7, R:40.3)
NP-coreference, generic NPs	PDT 2.0	1.8 (P:20, R:0.9)
bridging relations	PDT 2.0	0
Идентификация анафорического невыраженного субъекта, rule-based	PCEDT 2.0	61.5
Идентификация анафорического невыраженного субъекта, rule-based, с использованием английских текстов	PCEDT 2.0	69.5

Инструмент разметки синтаксических деревьев TrEd

* PML (Prague MarkUp Language) - XML-based format, разработанный для разметки трибанков

* настраиваемый tree editor TrEd (Pajas & Štěpánek 2008)

* расширения, включенные в инструмент как модули

* расширения для кореферентности и бриджинга, для дискурсивной разметки, для АЧ и т.д.

TrEd - coreference and bridging module

* **pre-**annotation данных с высокой вероятностью кореферентности
* supporting features включенных в TrEd – помощь **в процессе** ручной разметки

Pre-annotation

список пар выражений с высокой вероятностью кореферентности (ок. 6,000 пар)

Praha (сущ.) – *pražský* (прил.) *(Prague – Prague)*

He arrived in Prague and found the Prague atmosphere quite casual

Он приехал в Прагу, и пражская атмосфера ему очень понравилась

USA – United States of Amerika

Аннотирование

* *ручная pre-annotation узлов с идентичной t_lemma*

(1) Generál Jiří Nekvasil: V české armádě se hrozné věci nedějí
(7) Náčelník generálního štábu Jiří Nekvasil nám k tomu řekl, že nepořádek v armádě nikdy nezastíral, popírá však, že se v ní dějí hrozné věci.
(8) Poměry v útvarech podle něj odpovídají atmosféře ve společnosti.
(9) Armáda jde krok za krokem k lepšímu, říká Nekvasil.
(10) Za nepořádky náčelník generálního štábu považuje nízkou kázeň vojáků na veřejnosti, při výcviku nebo strážní službě.
(11) Starosti mu dělá nedodržování bezpečnostních zásad a neoprávněná manipulace se zbraněmi.
(14) Nekázeň se podle generála týká útvarů, jež transformace teprve zasáhne, zejména když mají být zrušeny.
(17) O panujícím strachu v armádě Nekvasil neví.
(18) Po mnoha letech podle něj dochází k narovnání vztahů mezi veliteli a podřízenými.
(19) Generál kromě toho připravuje nařízení, podle něhož se na něj budou moci obrátit všichni, kteří se domnívají, že se jim děje bezpráví.
(21) I proto se Nekvasil zabývá jednotlivými případy mladých důstojníků, kteří se rozhodli odejít do zálohy.
(22) Šéf generálního štábu považuje lustrace za uzavřenou záležitost.
(26) Jestliže by při jejich penzionování porušila zákon, soud by podle Nekvasila určitě prohrála.

Аннотирование

*ручная pre-annotation узлов с идентичной t_lemma

(1) Generál Jiří **Nekvasil**: V české armádě se hrozné věci nedějí
(7) Náčelník generálního štábu Jiří **Nekvasil** nám k tomu řekl, že nepořádek v armádě nikdy nezastíral, popírá však, že se v ní dějí hrozné věci.
(8) Poměry v útvarech podle něj odpovídají atmosféře ve společnosti.
(9) Armáda jde krok za krokem k lepšímu, říká **Nekvasil**.
(10) Za nepořádky náčelník generálního štábu považuje nízkou kázeň vojáků na veřejnosti, při výcviku nebo strážní službě.
(11) Starosti mu dělá nedodržování bezpečnostních zásad a neoprávněná manipulace se zbraněmi.
(14) Nekázeň se podle generála týká útvarů, jež transformace teprve zasáhne, zejména když mají být zrušeny.
(17) O panujícím strachu v armádě **Nekvasil** neví.
(18) Po mnoha letech podle něj dochází k narovnání vztahů mezi veliteli a podřízenými.
(19) Generál kromě toho připravuje nařízení, podle něhož se na něj budou moci obrátit všichni, kteří se domnívají, že se jim děje bezpráví.
(21) I proto se **Nekvasil** zabývá jednotlivými případy mladých důstojníků, kteří se rozhodli odejít do zálohy.
(22) Šéf generálního štábu považuje lustrace za uzavřenou záležitost.
(26) Jestliže by při jejich penzionování porušila zákon, soud by podle **Nekvasila** určitě prohrála.

Аннотирование

(1) Generál Jiří Nekvasil: V české armádě se hrozné věci nedějí
(7) Náčelník generálního štábu Jiří Nekvasil nám k tomu řekl, že nepořádek v armádě nikdy nezastíral, popírá však, že se v ní dějí hrozné věci.
(8) Poměry v útvarech podle něj odpovídají atmosféře ve společnosti.
(9) Armáda jde krok za krokem k lepšímu, říká Nekvasil.
(10) Za nepořádky náčelník generálního štábu považuje nízkou kázeň vojáků na veřejnosti, při výcviku nebo strážní službě.
(11) Starosti mu dělá nedodržování bezpečnostních zásad a neoprávněná manipulace se zbraněmi.
(14) Nekázeň se podle generála týká útvarů, jež transformace teprve zasáhne, zejména když mají být zrušeny.
(17) O panujícím strachu v armádě Nekvasil neví.
(18) Po mnoha letech podle něj dochází k narovnání vztahů mezi veliteli a podřízenými.
(19) Generál kromě toho připravuje nařízení, podle něhož se na něj budou moci obrátit všichni, kteří se domnívají, že se jim děje bezpráví.
(21) I proto se Nekvasil zabývá jednotlivými případy mladých důstojníků, kteří se rozhodli odejít do zálohy.
(22) Šéf generálního štábu považuje lustrace za uzavřenou záležitost.
(26) Jestliže by při jejich penzionování porušila zákon, soud by podle Nekvasila určitě prohrála.

TrEd ver. 1.4531 Tree Queries(1/1): http://euler.ms.mff.cuni.cz:8086/data/train-1/ln94200_169.t.gz

File Node Tree View Macros Setup Help Mode: PML_T_Bridging

Style: PML_T_Bridging

New query Import Suggest Connect Configure Edit query Edit node Edit subtree Filters Cut Copy Paste Paste new tree (Un)Expand (Un)Expand all

Add node NOT AND OR Equality Regexp Name Type Relation Add rel Optional

4/26

Generál Jiří Nekvasil: V české armádě se hrozné věci nedějí

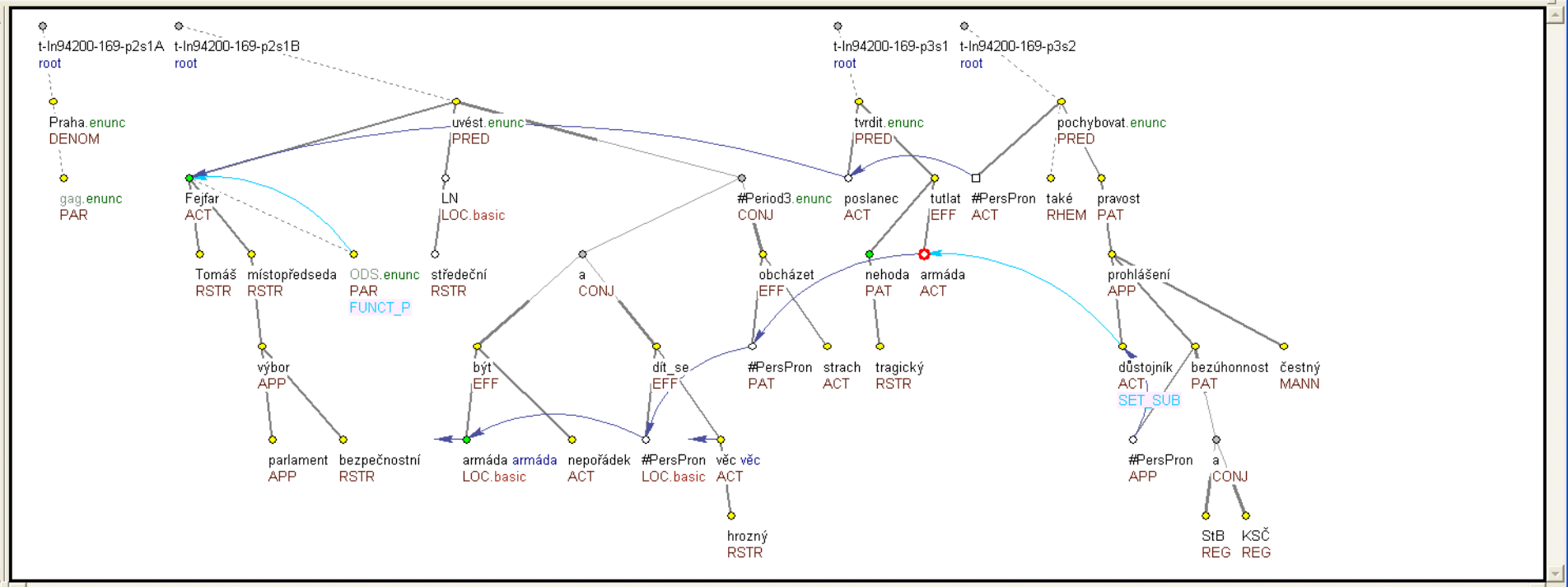
Praha (gag) -

V armádě je nepořádek a dějí se v ní hrozné věci... obchází ji strach, uvedl ve středečních LN místopředseda bezpečnostního výboru parlamentu Tomáš Fejfar (ODS).

--> Poslanec tvrdí, že tragické nehody armáda tutlá.
Pochybuje také o pravosti čestných prohlášení důstojníků o jejich bezúhonnosti ohledně StB a KSČ.
Tito lidé podle něj obklíčili armádu a vyřizují si při snižování počtů účty s bezúhonnými veliteli.

Náčelník generálního štábu Jiří Nekvasil nám k tomu řekl, že nepořádek v armádě nikdy nezastíral, popírá však, že se v ní dějí hrozné věci.
Poměry v útvarech podle něj odpovídají atmosféře ve společnosti.
Armáda jde krok za krokem k lepšímu, říká Nekvasil.
Za nepořádky náčelník generálního štábu považuje nízkou kázeň vojáků na veřejnosti, při výcviku nebo strážní službě.
Starosti mu dělá nedodržování bezpečnostních zásad a neoprávněná manipulace se zbraněmi.
Jen letos zahynulo k 11. červenci ve službě 6 základáků (většinou ve strážní službě) a 1 profesionál.
V době volna či dovolených 19 základáků (zpravidla dopravní nehody) a 8 profesionálů (sebevraždy).
Nekázeň se podle generála týká útvarů, jež transformace teprve zasáhne, zejména když mají být zrušeny.
Nejistota z budoucnosti prý snižuje svědomitost velitelů.
Jednotky, které se již stabilizovaly, odvádějí naopak dobré výsledky.

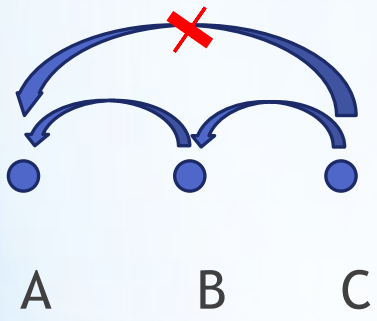
O panujícím strachu v armádě Nekvasil neví.
Po mnoha letech podle něj dochází k narovnání vztahů mezi veliteli a podřízenými.
Generál kromě toho připravuje nařízení, podle něhož se na něj budou moci obrátit všichni, kteří se domnívají, že se jim děje bezpráví.



Аннотирование

* поиск ближайшего антецедента

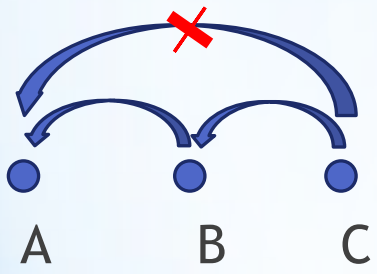
  автоматическое перенаправление стрелки на новую созданную стрелку к ближайшему антецеденту



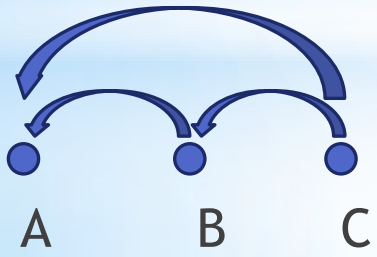
Аннотирование

* finding the nearest antecedent

  автоматическое перенаправление стрелки на новую созданную стрелку к ближайшему антецеденту



* автоматическое перенаправление в процессе разметки

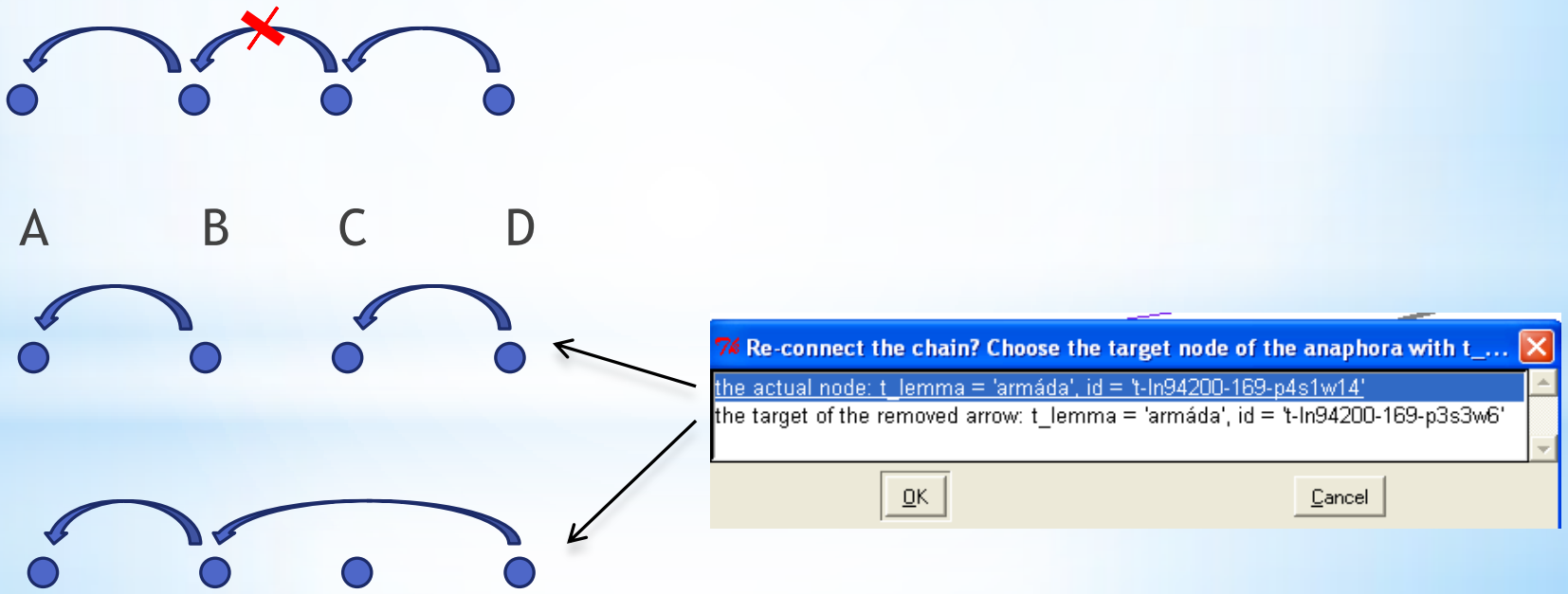


Аннотирование

* сохранение кореферентной цепочки

  разметчик стирает стрелку и нарушает цепочку:

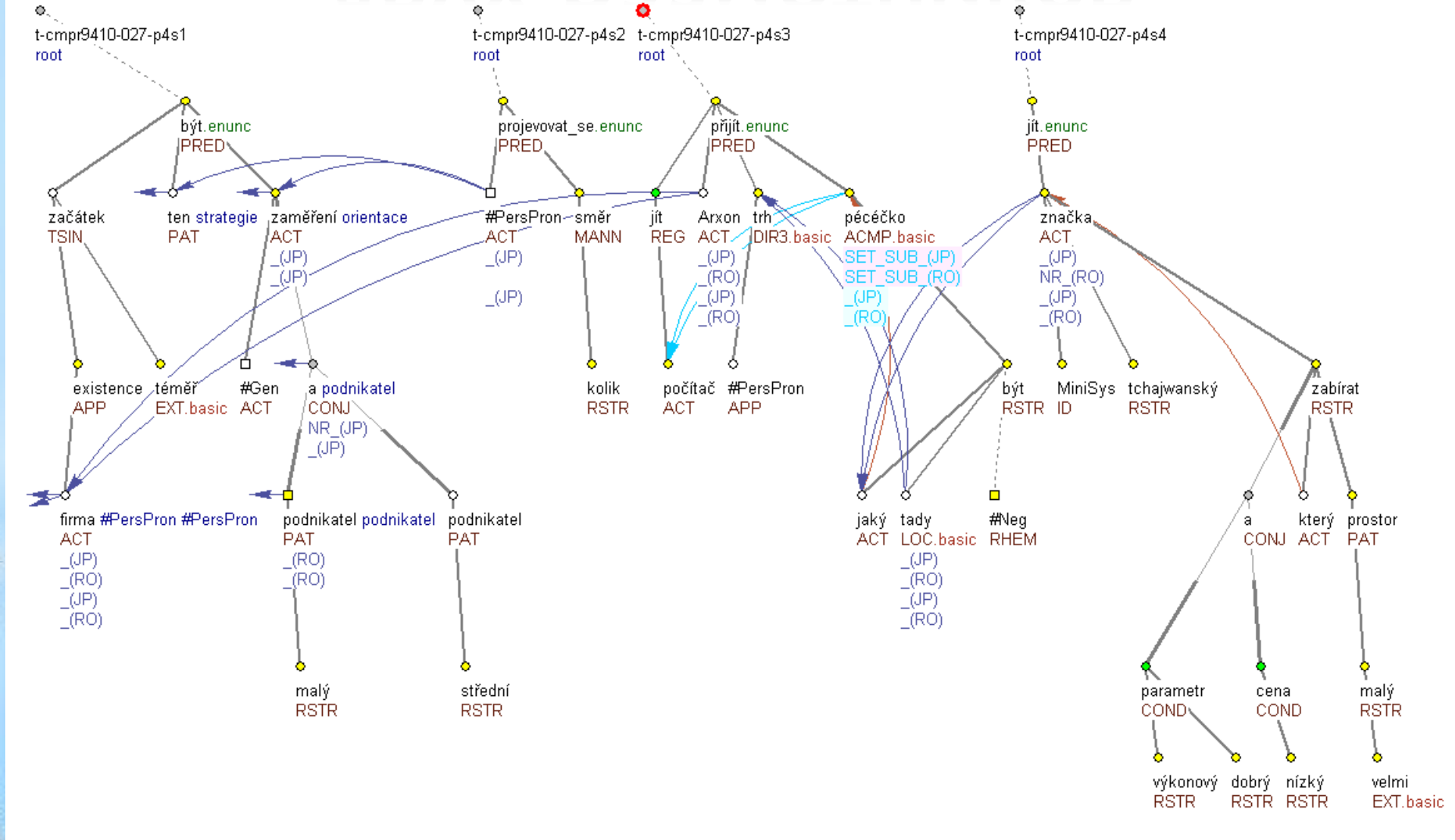
  TrEd восстанавливает нарушенную цепочку



Сравнение разметок

* визуализация параллельных разметок нескольких разметчиков
* используется для подсчета inter-annotator agreement

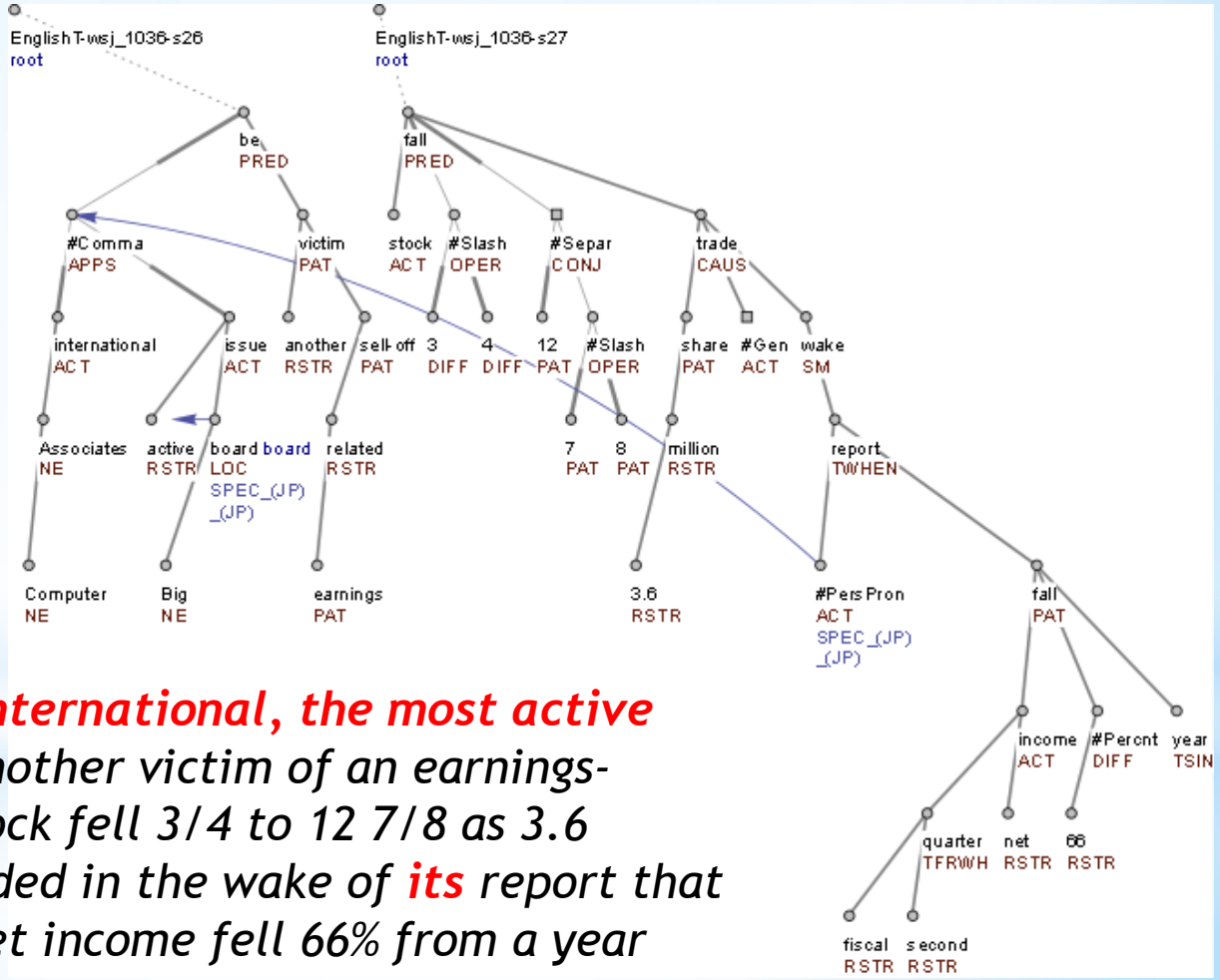
Сравнение разметок двух разметчиков



Преимущества разметки на синтаксических деревьях

- выбор маркабул автоматический (MIN-ID, максимальные группы)
- восстановленные нули
  - #Perspron: личные или поссессивные местоимения
  - #Cor: в контролирующих конструкциях
  - #Qcor: в квазиконтролирующих конструкциях, *He offered Jan {#QCor} protection.*
  - #Rcp: реципрок, *Мальчики поцеловались {#Rcp.PAT}.*
  - копирование опущенного узла
- нереферентные выражения:
  - аппозиции
  - сочинительные конструкции
  - двойные глагольные дополнения

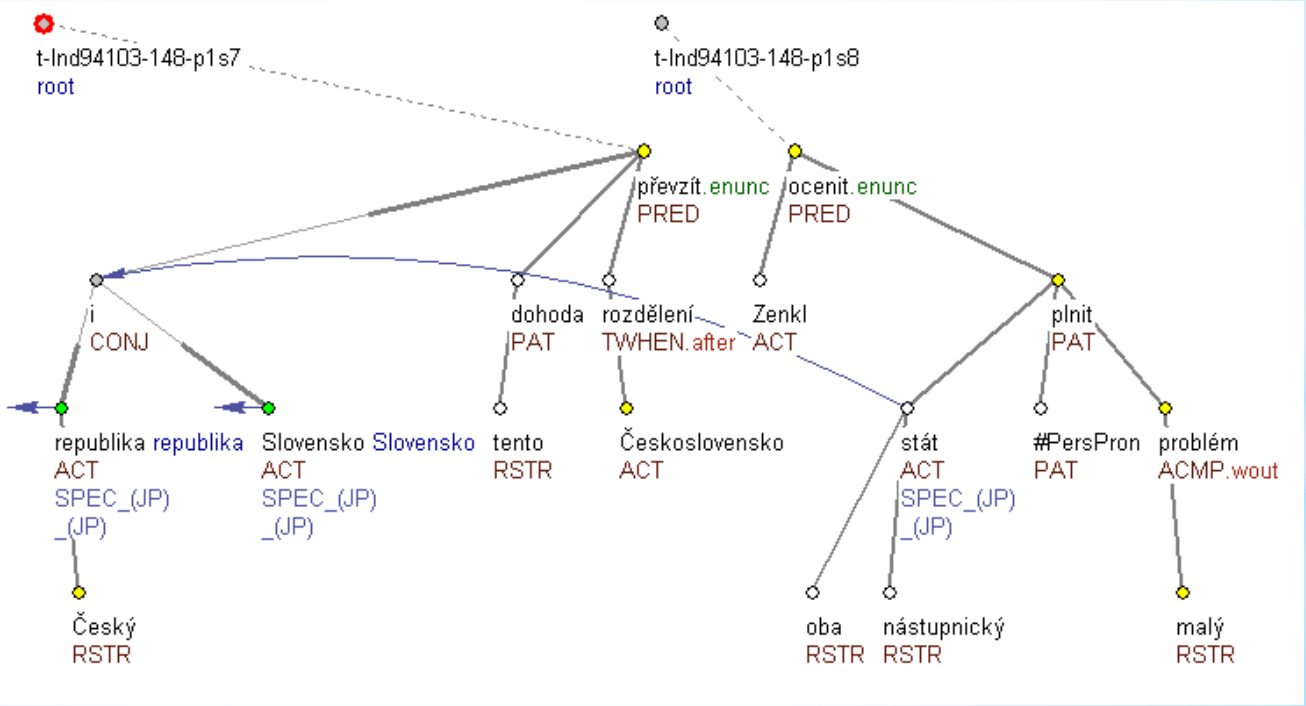
*Преимущества деревьев - Аппозиции



***Computer Associates International, the most active Big Board issue**, was another victim of an earnings-related sell-off. The stock fell 3/4 to 12 7/8 as 3.6 million shares were traded in the wake of **its** report that fiscal second-quarter net income fell 66% from a year ago.*

*Преимущества деревьев - сочинительные конструкции

Česká republika i Slovensko tuto dohodu po rozdělení Československa převzaly. Zenkl ocenil, že ji oba nástupnické státy plní bez nejmenších problémů. - Czech Republic and Slovakia took over this agreement after the split of Czechoslovakia. Zenkl appreciated that both successor states follow it without any problems.



Проблемы деревьев: предложные группы

* на глубинно-синтаксическом уровне предлоги включены в лемму узла, которым они управляют
* PP размечаются как NP (*near Prague* = *Prague*, *before the war - during the war - after the war*)

неидеальное, но технически удобное решение? (иначе очень низкий agreement для *in Prague - about Prague - for Prague* vs. *in Prague - around Prague - above Prague*) --- не всегда понятно, что есть что:

Zatím se posunuje stále více za Prahu, čímž ztrácí na své účelnosti z hlediska dopravních spojení do jednotlivých částí města Na druhé straně by tu asi mohlo být víc pozemků vhodných k podnikání. Po dálnici bychom se měli svézt z Prahy až do Českých Budějovic, v roce 1997 pravděpodobně projedou první vozidla po dálnici Praha -Plzeň, dokončena by měla být i dálnice D8 z Prahy do Ústí nad Labem.

(= So far, people begin to move away from Prague, ... various parts of the city. On the other hand, there could be more lands suitable for business there. Highways could take us from Prague up to CeskéBudejovice … .)

*Вопросы?

Дискурсивная разметка

коннекторы, аргументы, деревья

Обзор существующих дискурсивных разметок

- **Древесные структуры**
  - Rhetorical Structure Theory (RST- www.sfu.ca/rst; Carlson и др., 2001; сейчас Maite Taboada, 2006-2013; …)
  - The Potsdam Commentary Corpus (Stede et al. 2004)
  - Segmented Discourse Representation Theory (SDRT, Asher & Lascarides, 2003): AnnoDis (Afantenos и др. 2012), Ascher – STAC corpus
  - “Veins theory” (VT), Cristea и др., 1998
- **Недревесные структуры**
  - Discourse Graphbank: Wolf - Gibson (2005)
- **Лексический подход** (выделение коннекторов)
  - Penn Discourse TreeBank (Joshi и др. 2005-8),
  - Hindi Discourse Bank (Joshi и др.)
  - Prague Dependency Treebank (Poláková и др. 2014)
  - German-English Contrasts in Cohesion (Kunz et al. 2014)

* Rhetorical Structure Theory (RST) Treebank (Carlson, Marcu 2001)

[Still, analysts don't expect the buy -back to significantly affect per -share earnings in the short term.][16] ["The impact won't be that great,"] [17] [said Graeme Lidgerwood of First Boston Corp.] [18] [This is in part because of the effect] [19] [of having to average the number of shares outstanding,] [20] [she said.][21] [In addition,] [22] [Mrs. Lidgerwood said,] [23] [Norfolk is likely to draw down its cash initially] [24] [to finance the purchases][25] [and thus forfeit some interest income.][26] wsj_1111

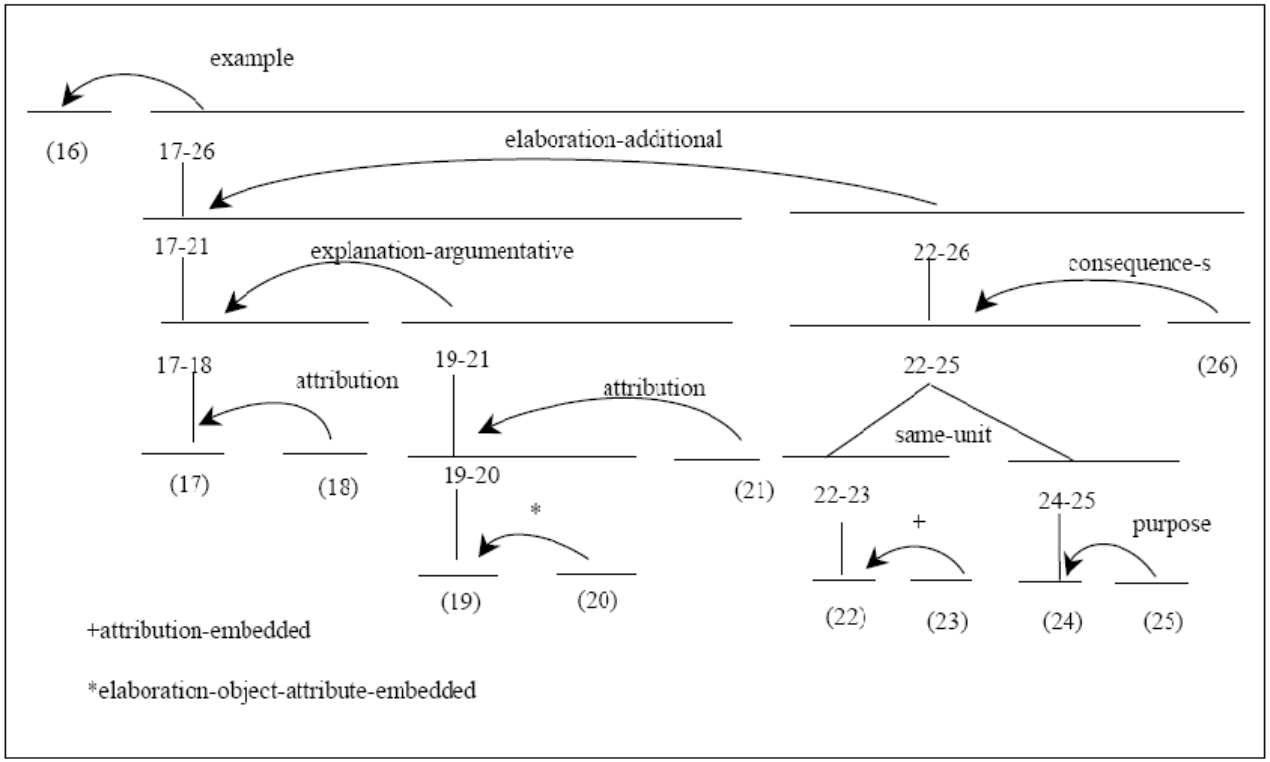
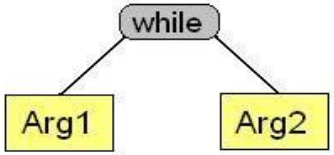


Figure 1. Discourse sub-tree for multiple sentences

*Penn Discourse TreeBank 2.0

John eats porridge for breakfast, <u>while</u>
Mary eats muesli.



The asbestos fiber, crocidolite, is unusually resilient once it enters the lungs, with even brief exposures to it causing symptoms that show up decades later, researchers said.
Lorillard Inc., the unit of New York-based Loews Corp. that makes Kent cigarettes, stopped using crocidolite in its Micronite cigarette filters in 1956.

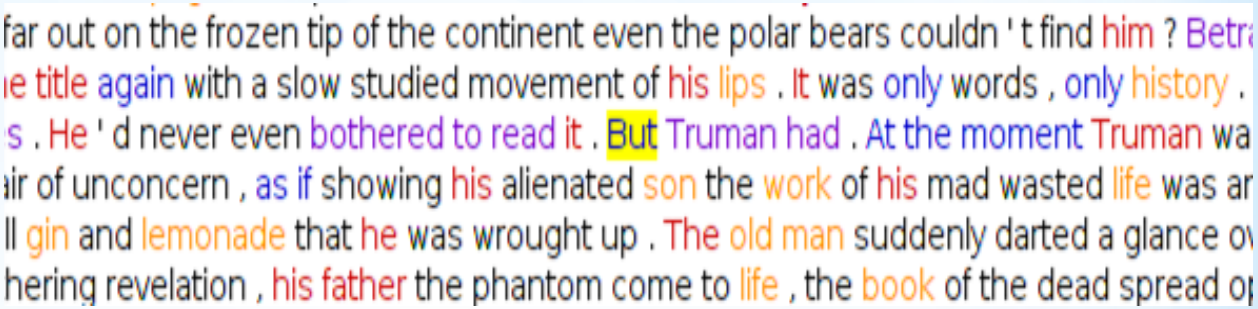
Although preliminary findings were reported more than a year ago, the latest results appear in today's New England Journal of Medicine, a forum likely to bring new attention to the problem.

A Lorillard spokewoman said, "This is an old story.
We're talking about years ago before anyone heard of asbestos having any questionable properties.
There is no asbestos in our products now."

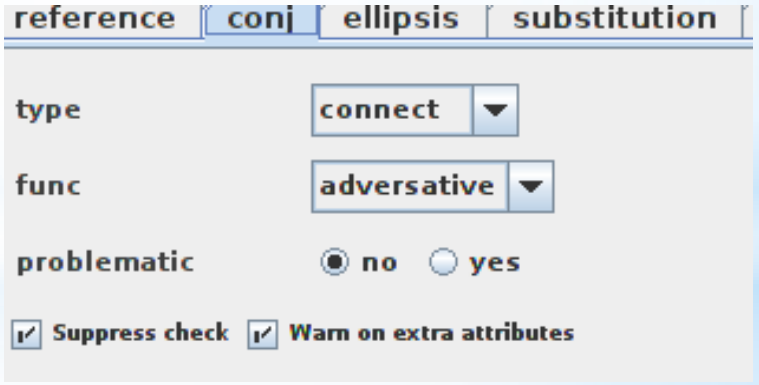
Neither Lorillard nor the researchers who studied the workers were aware of any research on smokers of the Kent cigarettes. "We have no useful information on whether users are at risk," said James A. Talcott of Boston's Dana-Farber Cancer Institute.

Conn/AltLex | Conn/AltLex Attr | Arg1 | Arg1 Attr | Arg2 | Arg2 Attr | Sup1 | Sup2

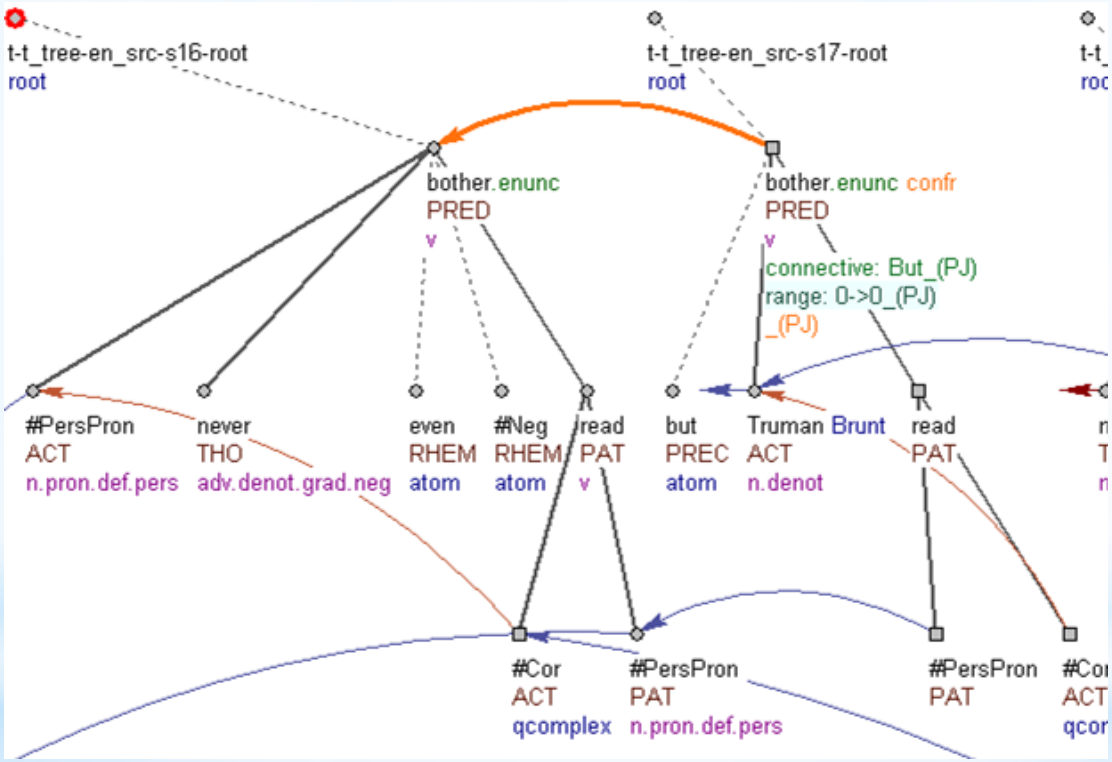
German-English Contrasts in Cohesion (Kunz et al. 2014)



He 'd never even bothered to read it. But Truman had.



Prague Dependency Treebank (Poláková et al. 2014)



He 'd never even bothered to read it. But **Truman had.**

* ***"Slovak elite was disappointed by the political choice of Slovakia. Therefore, the majority of quality specialists remained in Prague."***

Словацкая элита была разочарована политическим выбором Словакии. Поэтому большинство хороших специалистов осталось в Праге.

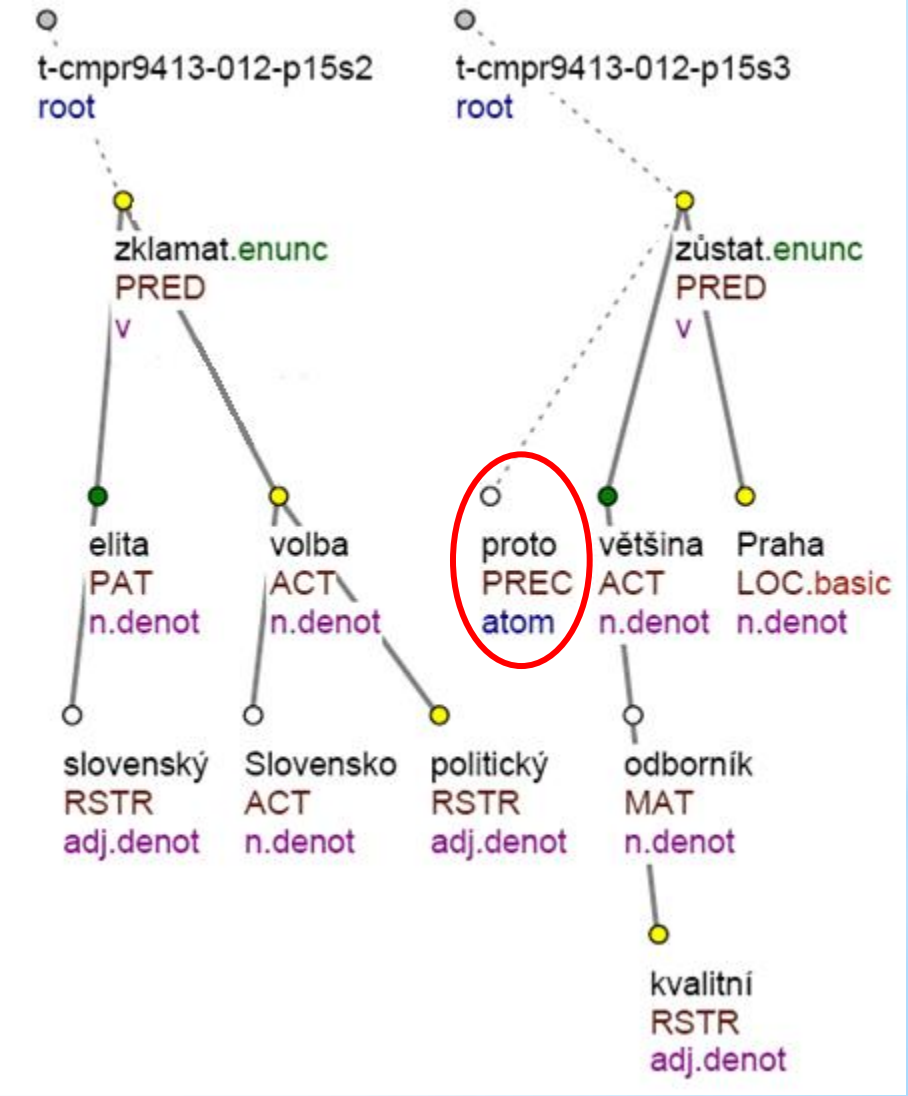
* ***"Slovak elite was disappointed by the political choice of Slovakia. Therefore, the majority of quality specialists remained in Prague."***

Словацкая элита была разочарована политическим выбором Словакии. Поэтому большинство хороших специалистов осталось в Праге.

* ***"Slovak elite was disappointed by the political choice of Slovakia. Therefore, the majority of quality specialists remained in Prague."***

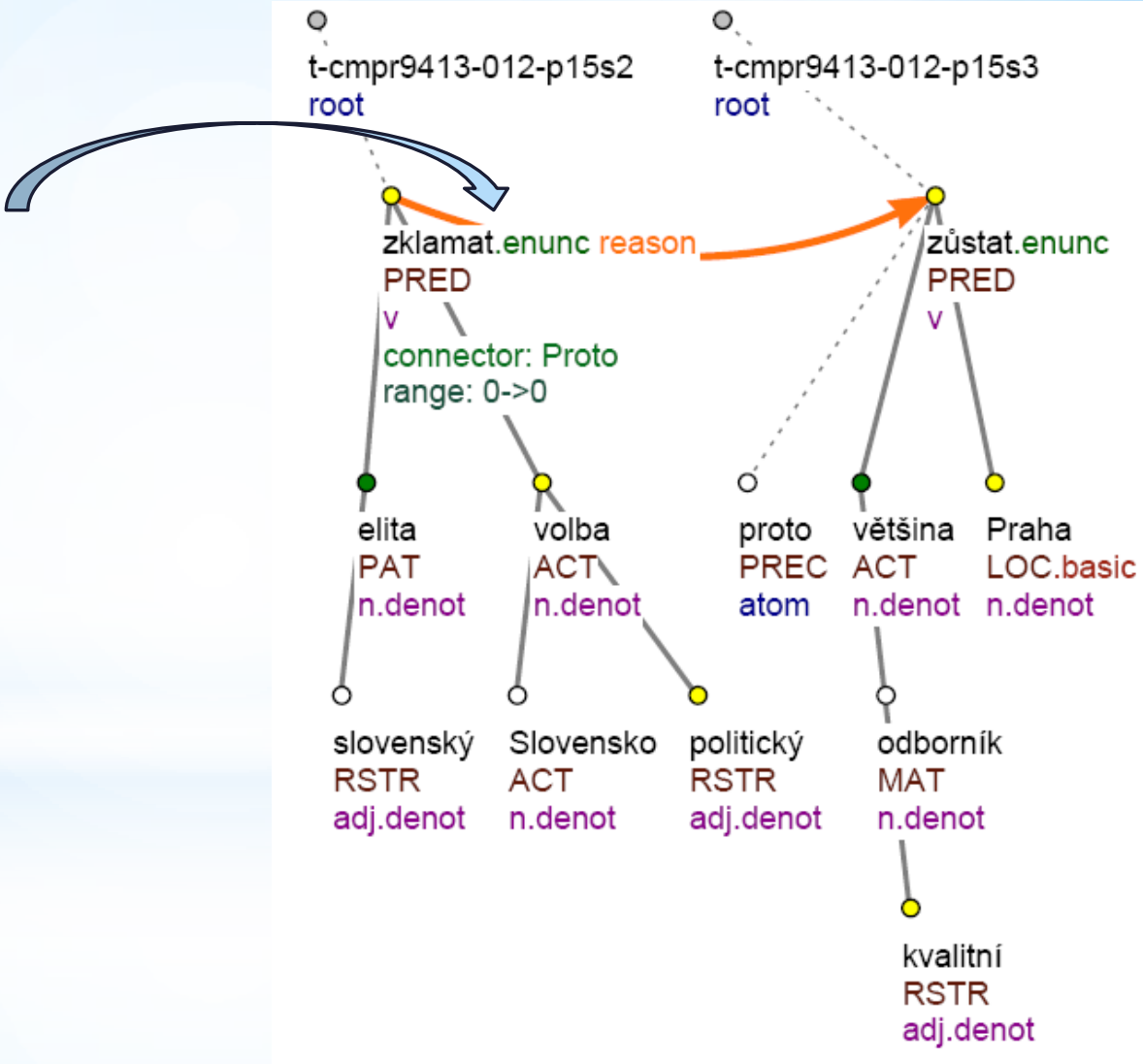
Словацкая элита была разочарована политическим выбором Словакии. *Поэтому большинство хороших специалистов осталось в Праге.*

* *"Slovak elite was disappointed by the political choice of Slovakia. **Therefore**, the majority of quality specialists remained in Prague."*



* *"Slovak elite was disappointed by the political choice of Slovakia. **Therefore**, the majority of quality specialists remained in Prague."*

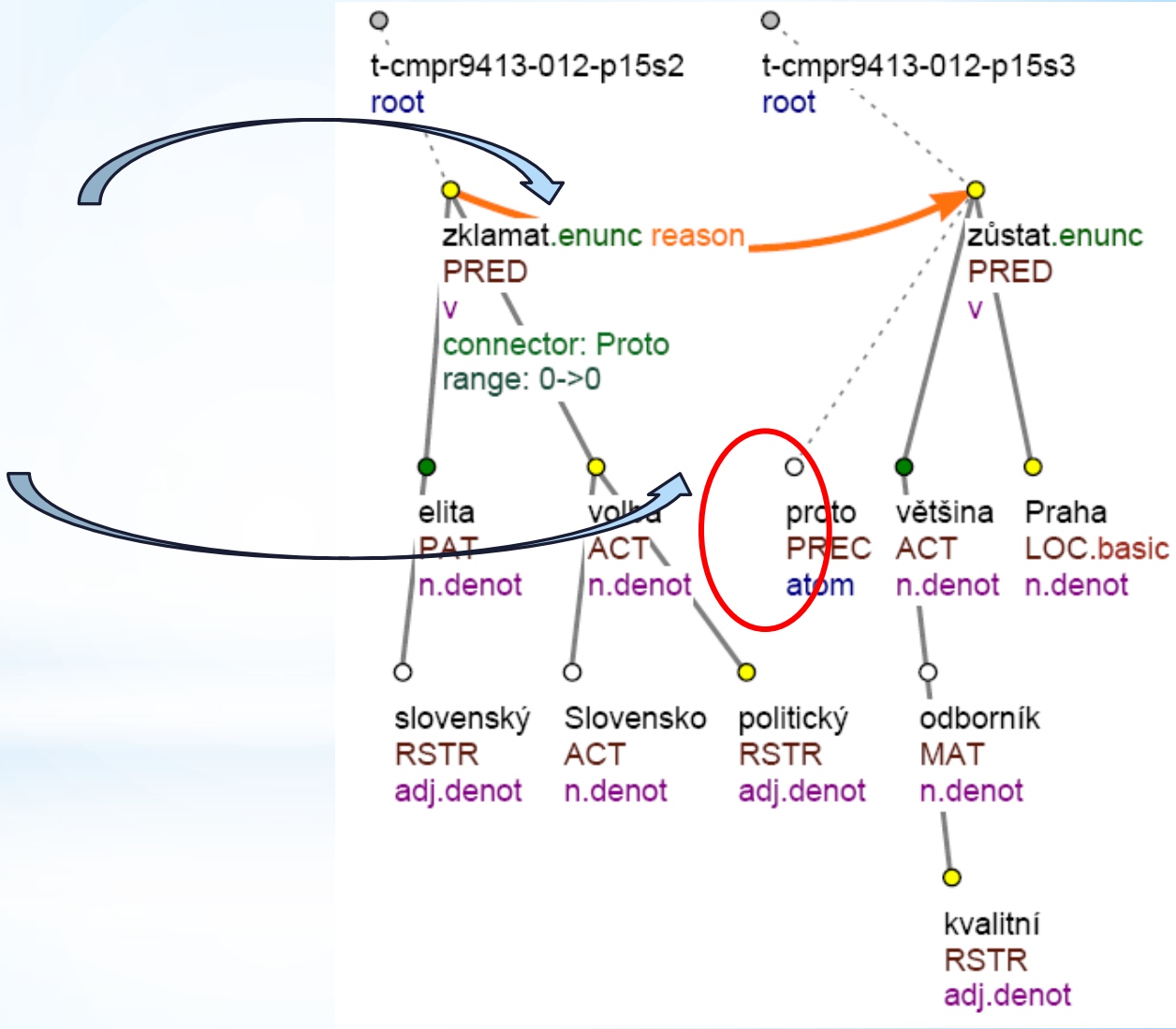
* аргументы связываются стрелкой



* *"Slovak elite was disappointed by the political choice of Slovakia. **Therefore**, the majority of quality specialists remained in Prague."*

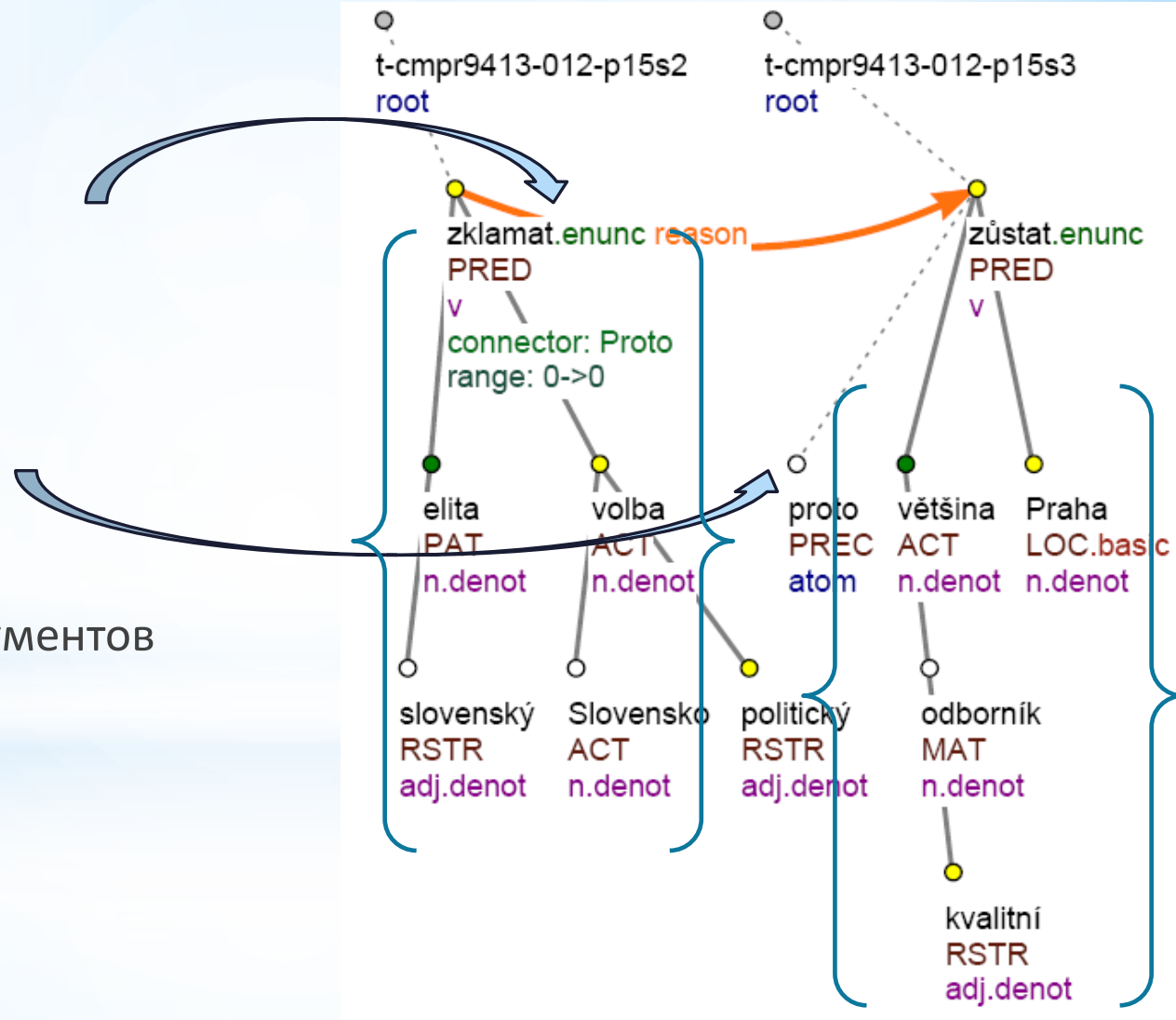
* аргументы связываются стрелкой

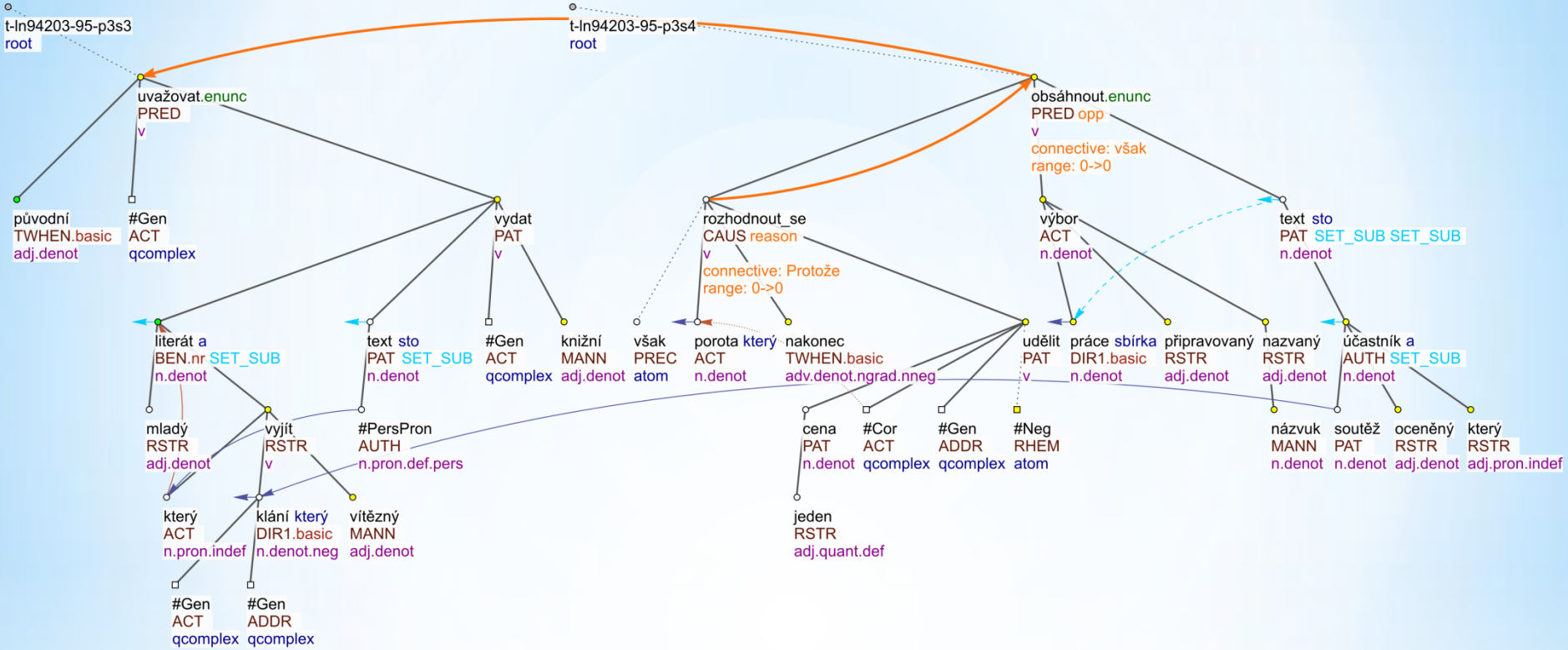
* добавляем коннектор



* *"Slovak elite was disappointed by the political choice of Slovakia. **Therefore**, the majority of quality specialists remained in Prague."*

* аргументы связываются стрелкой
* добавляем коннектор
* обозначаем размер аргументов





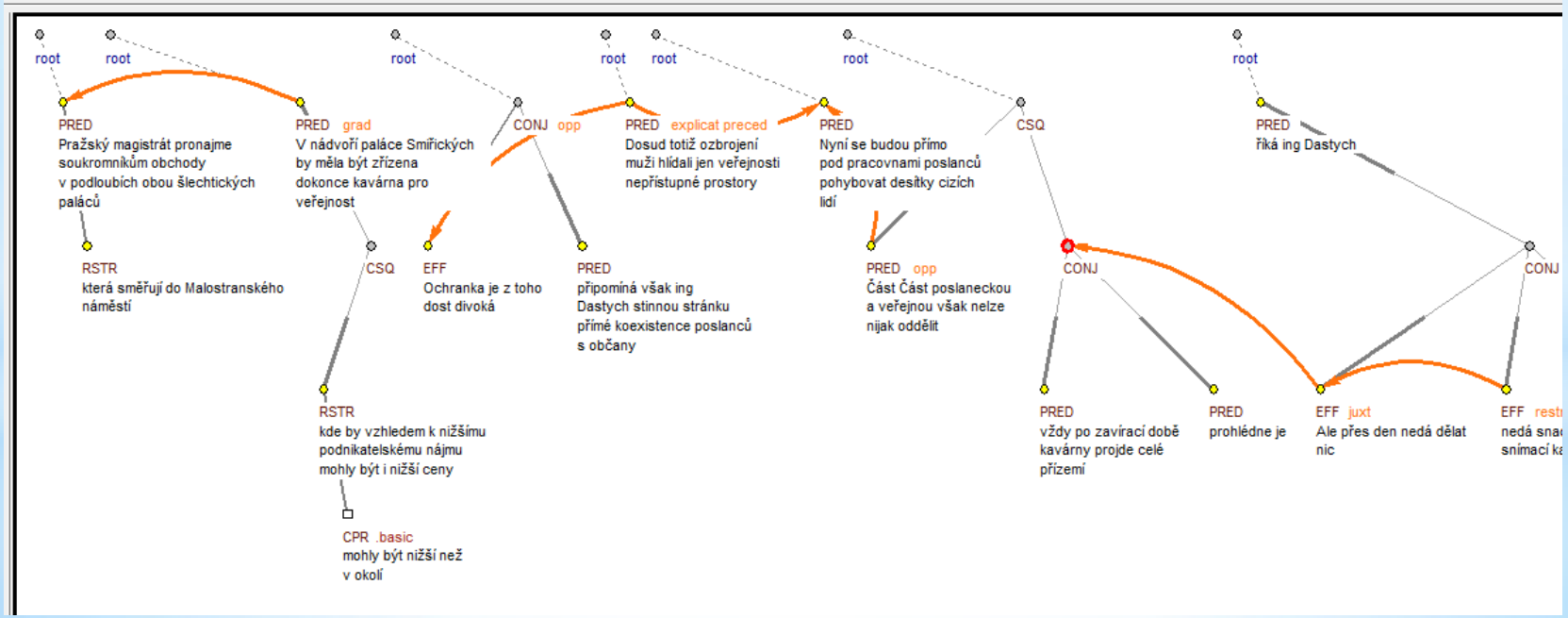
Původně se uvažovalo o tom, že mladému literátovi, jenž by vyšel z klání vítězně, budou jeho texty vydány knižně. Protože se však porota rozhodla nakonec první cenu neudělit, obsáhne připravovaný výbor z prací nazvaný Názvuky texty všech oceněných účastníků soutěže.

Initially, the book of the winner was supposed to be published. However, as the jury decided not to give the first price, the collected works of all authors will be published.

Визуализация в TrEd - Compact View

Ochranka je z toho dost divoká, připomíná však ing. Dastych stinnou stránku přímé koexistence poslanců s občany.
Dosud totiž ozbrojení muži hlídali jen veřejnosti nepřístupné prostory.
Nyní se budou přímo pod pracovnami poslanců pohybovat desítky cizích lidí.
--> Část poslaneckou a veřejnou však nelze nijak oddělit, proto vždy po zavírací době kavárny projde ochranka se psem celé přízemí a prohlédne je.
Ale přes den se nedá dělat nic, snad jen instalovat snímací kamery a služba je bude muset nějak ohlídat, říká ing. Dastych.

Tunel, nebo most?



Отношения в дискурсивной разметке

TEMPORAL	CONTINGENCY	COMPARISON (CONTRAST)	EXPANSION
precedence - succesion	reason - result	confrontation	conjunction
synchronous	pragmatic reason – result	opposition	instantiation
	purpose	pragmatic contrast	specification
	explication	restrictive opposition + exception	equivalence
	condition	concession	generalization
	pragmatic condition	correction (replacement)	conjunctive alternative
		gradation	disjunctive alternative

Temporal - подтипы

TEMPORAL - подтипы	Примеры
precedence – succession	*The lamp extinguished. **Before that**, it only sputtered for a while.*
synchronous	*The tenth hour struck **and** the lamp was still shining.*

CONTINGENCY – подтипы

обстоятельства, отношения имплицитного типа

CONTINGENCY - подтипы	Примеры
reason - result	*He was dismissed **because** he worked irresponsibly.*
pragmatic reason – result	*Grandmother is home **because** the lights are on in the kitchen.*
condition – result of the condition	*I will make pancakes. **But** first you must buy eggs.*
pragmatic condition – result of the condition	***If** you do understand it, so I do not.*
purpose	*She goes to train regularly. She wants to lose weight.*
explication	*He is a thief. He was shop lifting.*

COMPARISON – подтипы

COMPARISON - подтипы	Примеры
confrontation	*The worker is mortal, the work is alive, Anthony is dying, the bulb is singing.*
opposition	*He heard everything. **But** he saw nothing.*
pragmatic opposition	*It is going to rain this weekend. **But** Czechs will block the highways anyway*
restrictive opposition + exception	*I will come. I **only** do not know when*
concession (уступка)	*They died. **And yet** they still speak.*
correction (or replacement) + chosen alternative (substitution)	*He did not wait at home. He followed her to work.*
gradation	*He was running. What is more, he was speeding.*

EXPANSION – подтипы

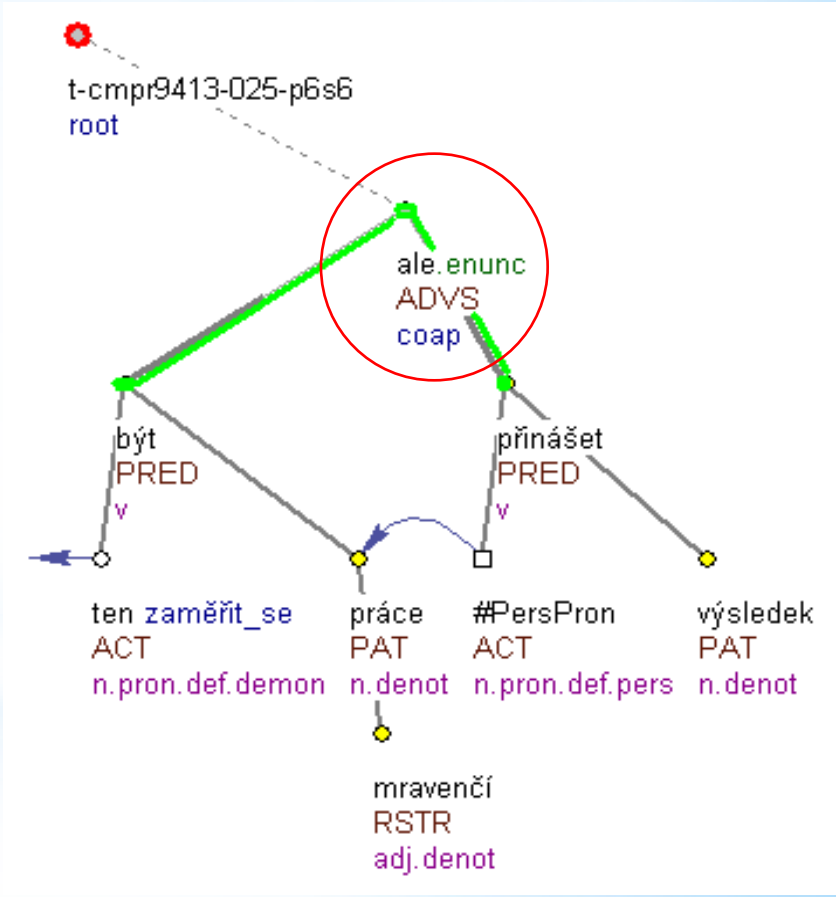
EXPANSION - подтипы	Примеры
conjunction	*He went straight. He did not look left nor right.*
instantiation	*She never spent evenings at home. **For example**, she went for walks with friends.*
specification	*He tries to reduce debt. He earns more money.*
equivalence	*The method is up to you. Just do it by yourself.*
generalization	*They lent him some money. **In short**, they helped him.*
conjunctive alternative	*We may go to the cinema. **Or** we may go for a coffee* (or both).
disjunctive alternative	*Behave decently. **Or** do not come here!*

Коннекторы (connectives)

- в рамках предложения (intra-sentential)
- за рамками предложения (inter-sentential)
- имплицитные коннекторы
- коннекторы х дискурсивные маркеры

Intra-sentential Discourse Connectives коннекторы в рамках предложения

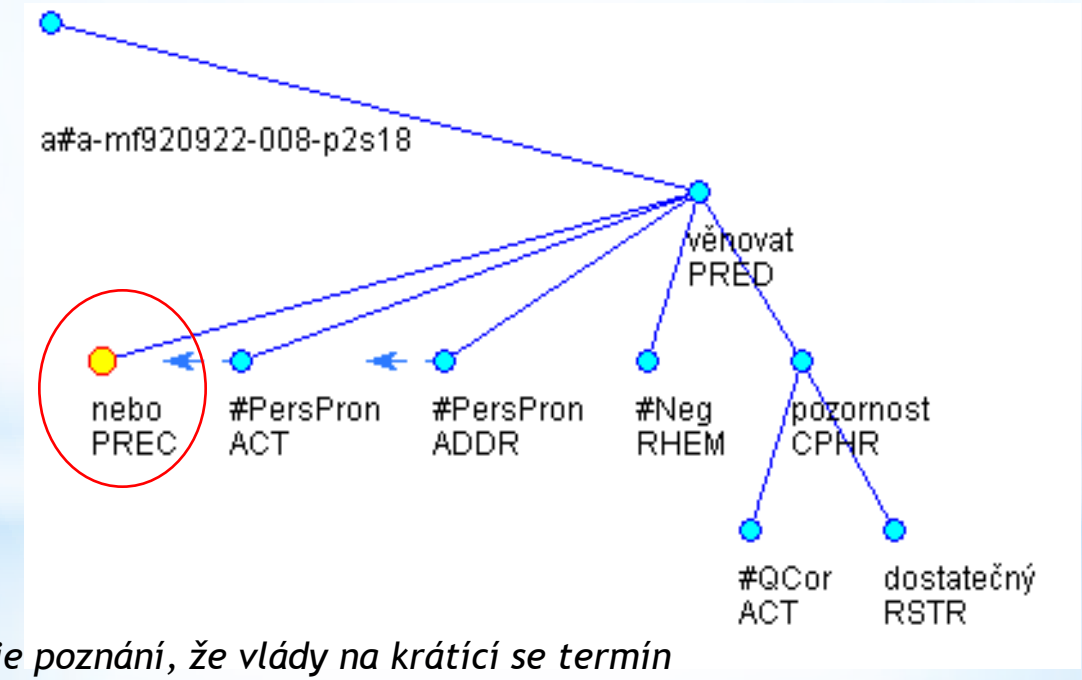
* Соединяет части дерева (предложения), часть синт. структуры
* сочинительные союзы (на картинке)
* или подчинительные клаузы в сложноподчиненных предложениях



Je to mravenčí práce, ale přináší výsledky.

* Inter-sentential Discourse Connectives коннекторы за рамками предложения

* обозначены функтором PREC на гл.-синт. уровне (отсылка к PREceding Context)



(Obzvlášť tristní je poznání, že vlády na krátící se termín blokace zákona o bankrotu zřejmě jednoduše zapomněly.)
Nebo ***mu nevěnovaly dostatečnou pozornost.***

AltLexes – отдельная разметка

* *Nikdy netrávila večery doma. Chodila <u>například</u> na procházky s přáteli.*

  Она никогда не сидела по вечерам дома. Ходила, <u>например</u>, гулять с друзьями

  Instantiation

* *Metoda záleží jenom na vás. <u>Prostě</u> to udělejte podle sebe.*

  Методика зависит от вас. <u>Просто</u> сделайте так, как считаете нужным.

  Equivalence

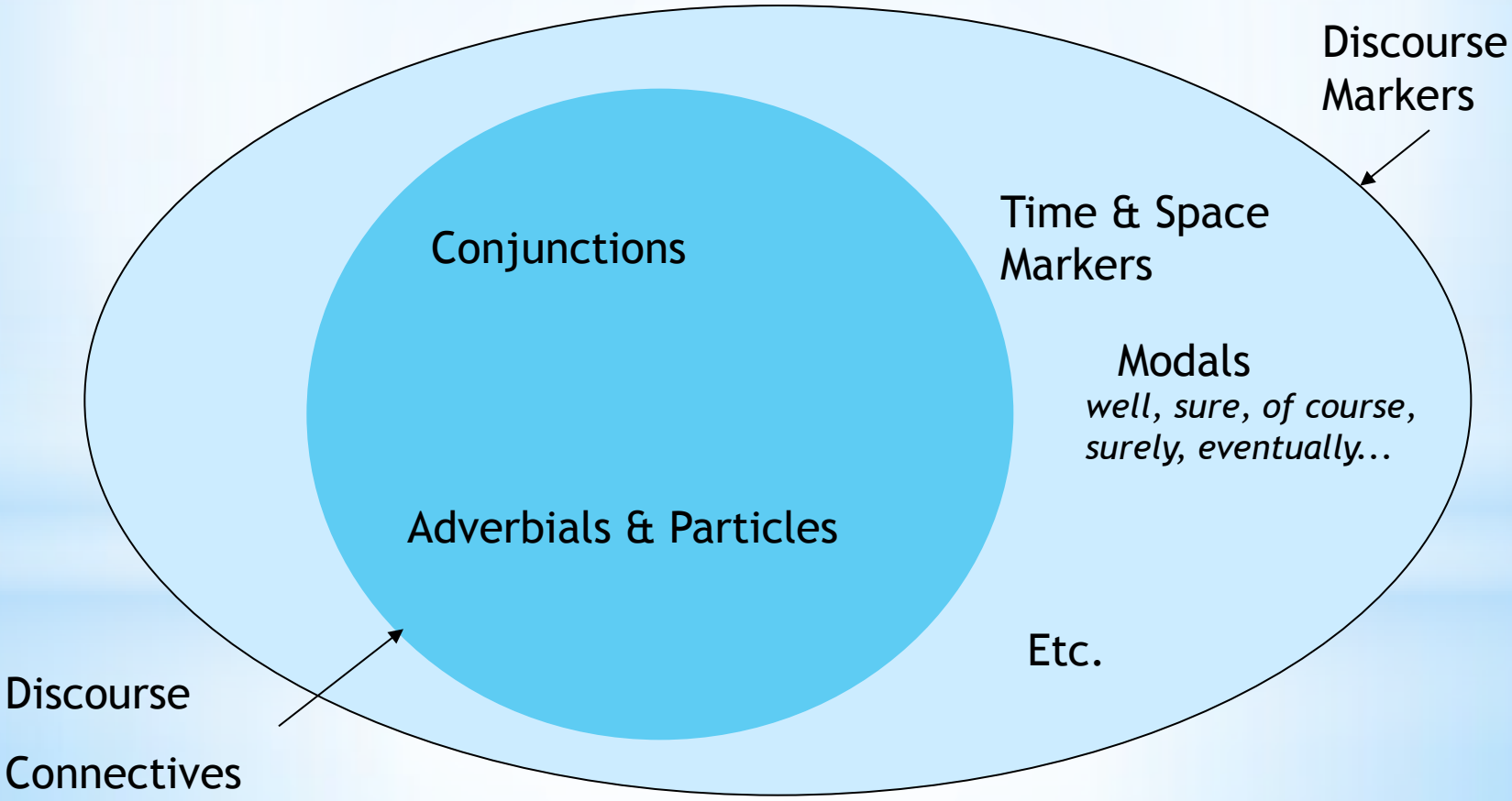
Пока не размечаются

- Отношения, связанные имплицитными коннекторами (не выражены в тексте)

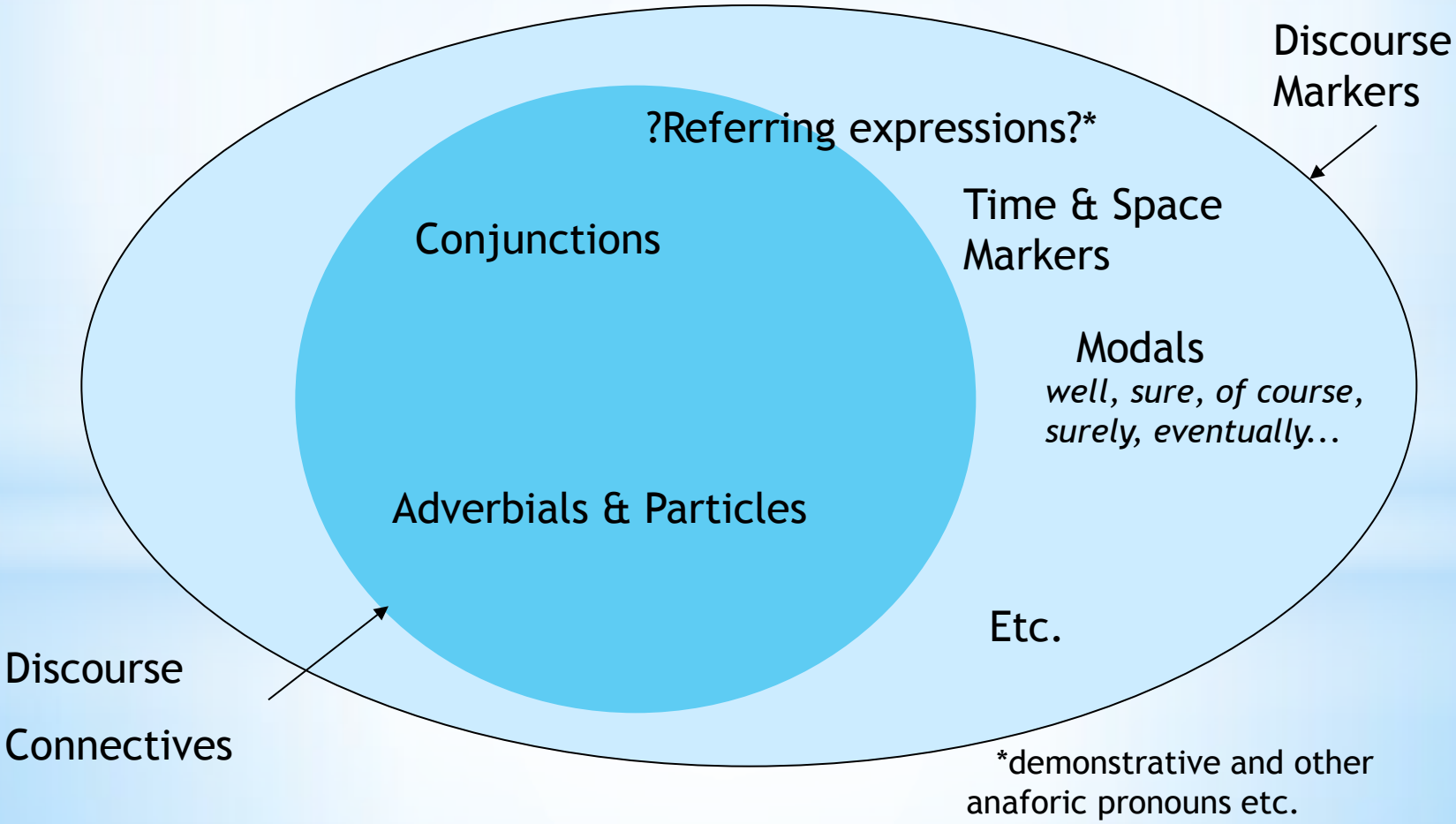
Jel do zatáček opatrně. Vždy si nadjížděl.

Specification

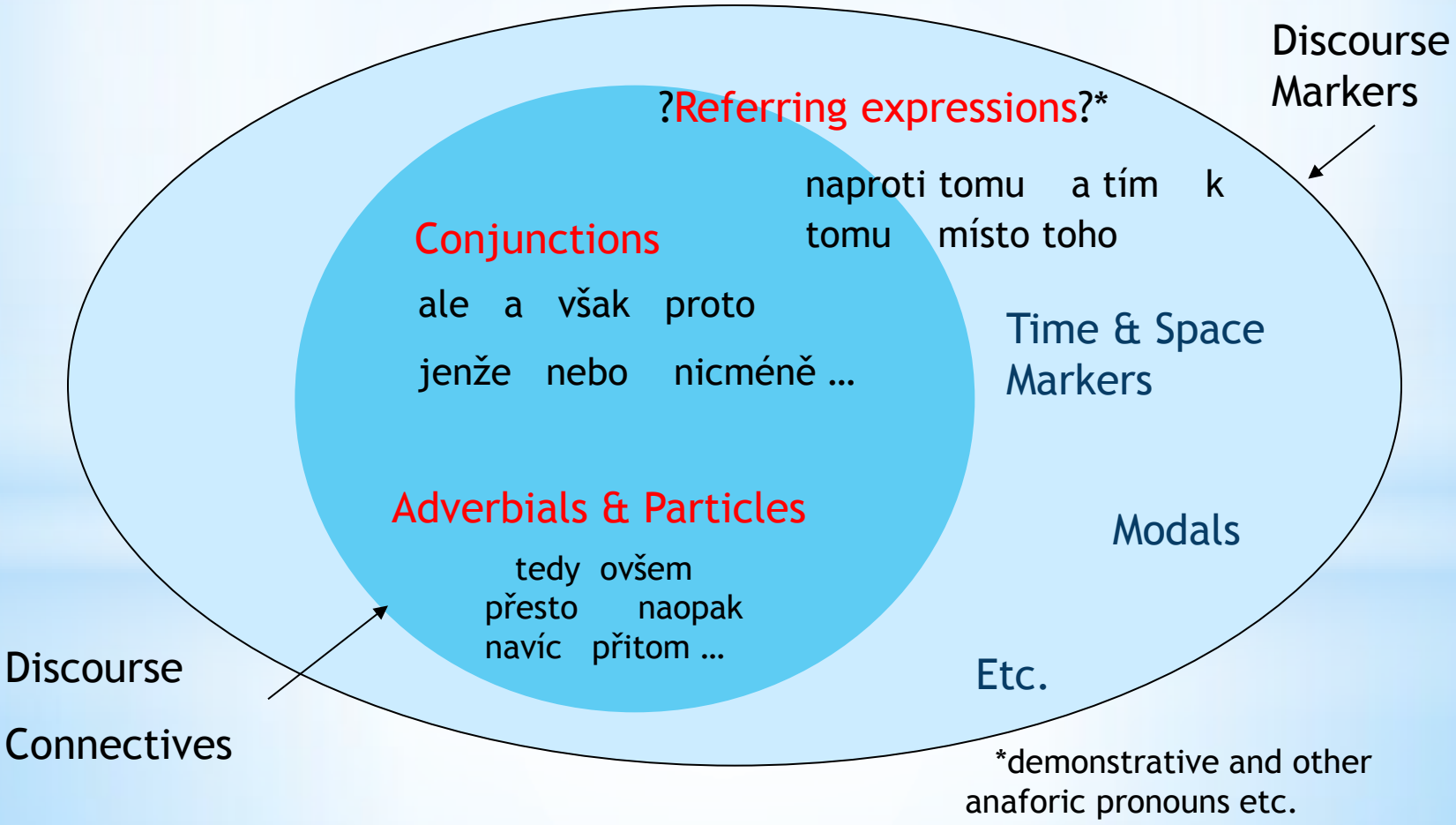
*Коннекторы х Дискурсивные маркеры



*Коннекторы х Дискурсивные маркеры



*Коннекторы x Дискурсивные маркеры



Свойства коннекторов

- Неизменяемая лексическая единица или фразеологизованное выражение
- Не является частью синтаксической структуры предложения
- Соединяет два сегмента текста
- Может иметь прагматические функции

John is home because the lights are on in the house.
Pragmatic cause

Многозначность коннекторов

- *Pršelo, ale deštník si nevzal.* Уступка

  Шел дождь, но зонт он не взял.

- *Nespal, ale vymýšlel plán na zítřek.* Противопоставление

  Он не спал, но придумывал план на завтра.

- *Nesportuji, ale na plovárnu si občas zajdu.* Исключение

  Спортом я не занимаюсь, но плавать иногда хожу

- *Dal si nejen hlavní jídlo, ale objednal si i zákusek.* Градация

  Он не только еду заказал, но и десерт.

- *To je ale krásně!* вообще не коннектор

  Вот это но прекрасно!

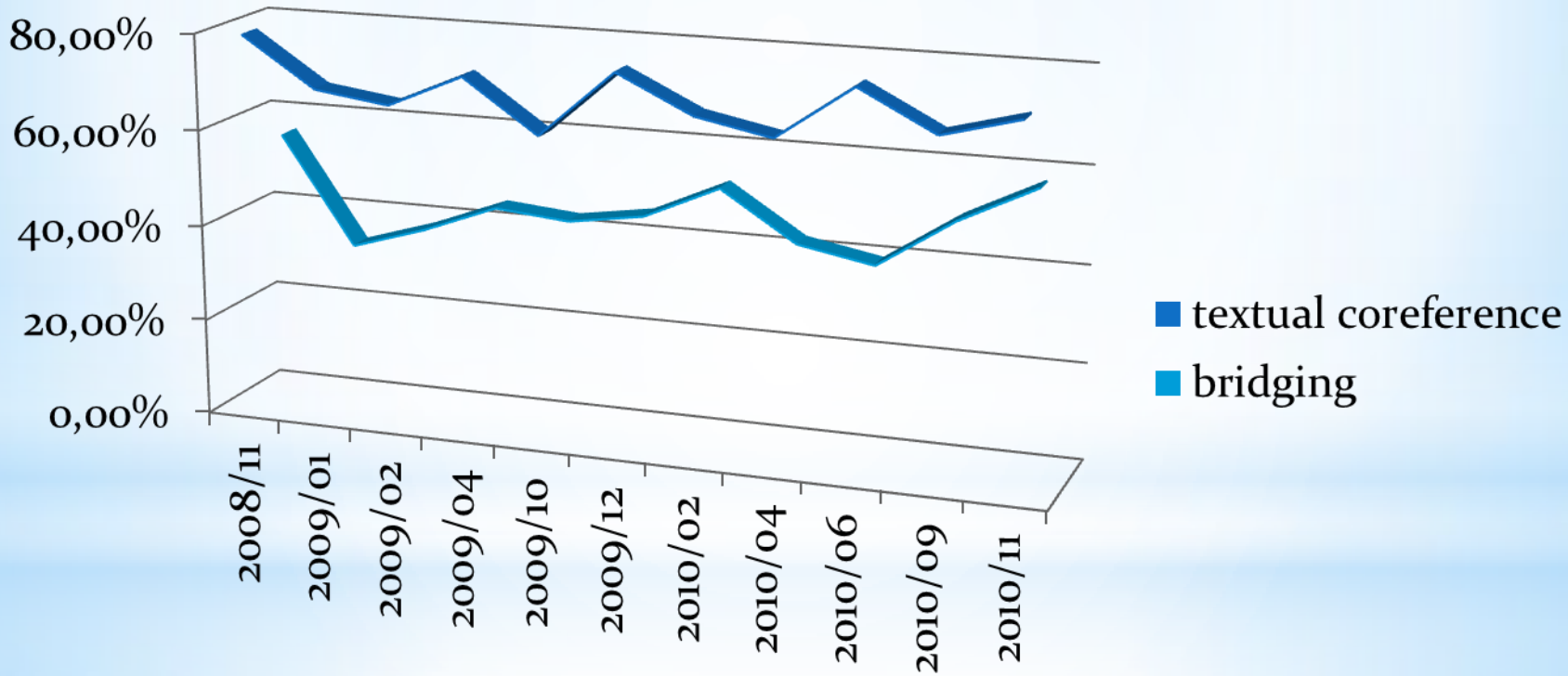
- *Byl teplý, ale zamračený den.* *коннектор, но не дискурсивный*

  Был теплый, но пасмурный день.

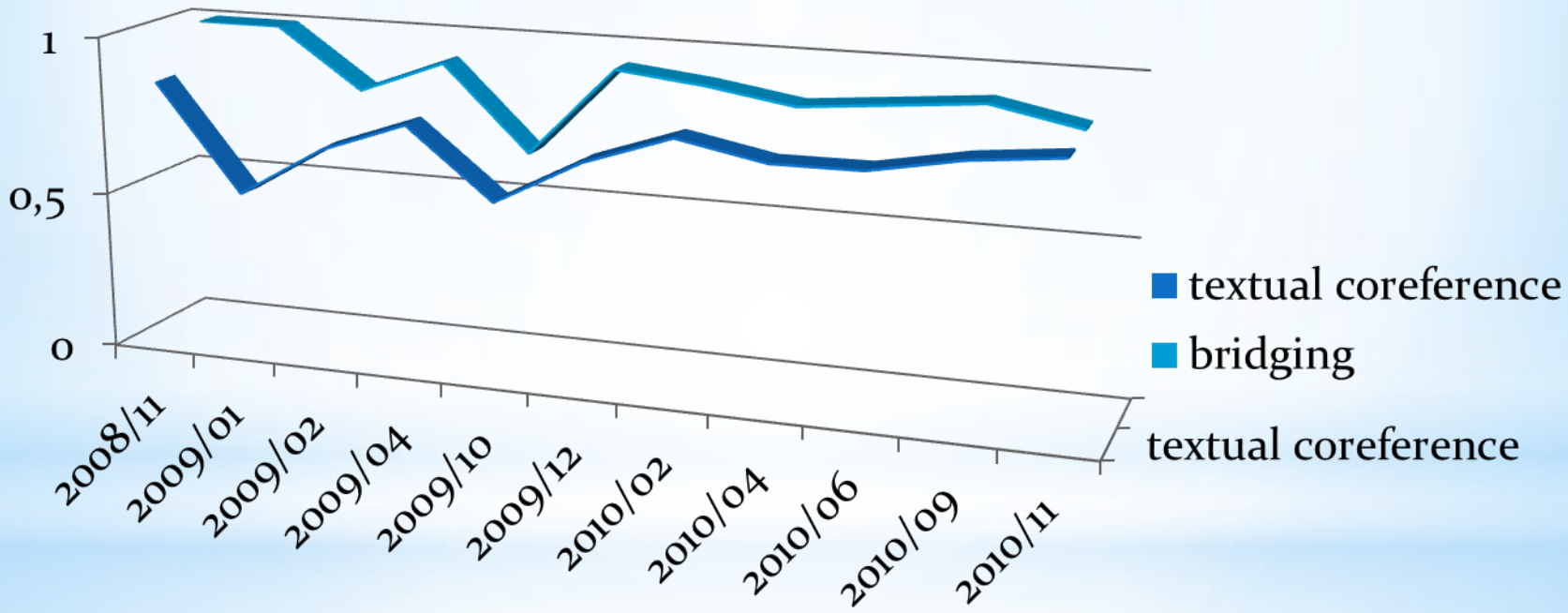
Спасибо за внимание!

http://ufal.mff.cuni.cz/discourse

*Inter-annotator Agreement



*Inter-annotator Agreement: *kappa* for Types



*Types of disagreement

* the relation should or should not be annotated for coreference/bridging
* what is the correct antecedent of a given noun phrase
* distinguishing between the bridging anaphora and the textual coreference
* selecting the type of the bridging anaphora or the textual coreference

Annotating / not annotating a relation

A když už byla knížka hotova, tak se zjistilo, že je praktická i pro rodiče. V této knize je poučení, jak snášejí děti rozvod a jak na něj reagují, a návod, jak se mají rodiče chovat, aby se utrpení dětí snížilo.

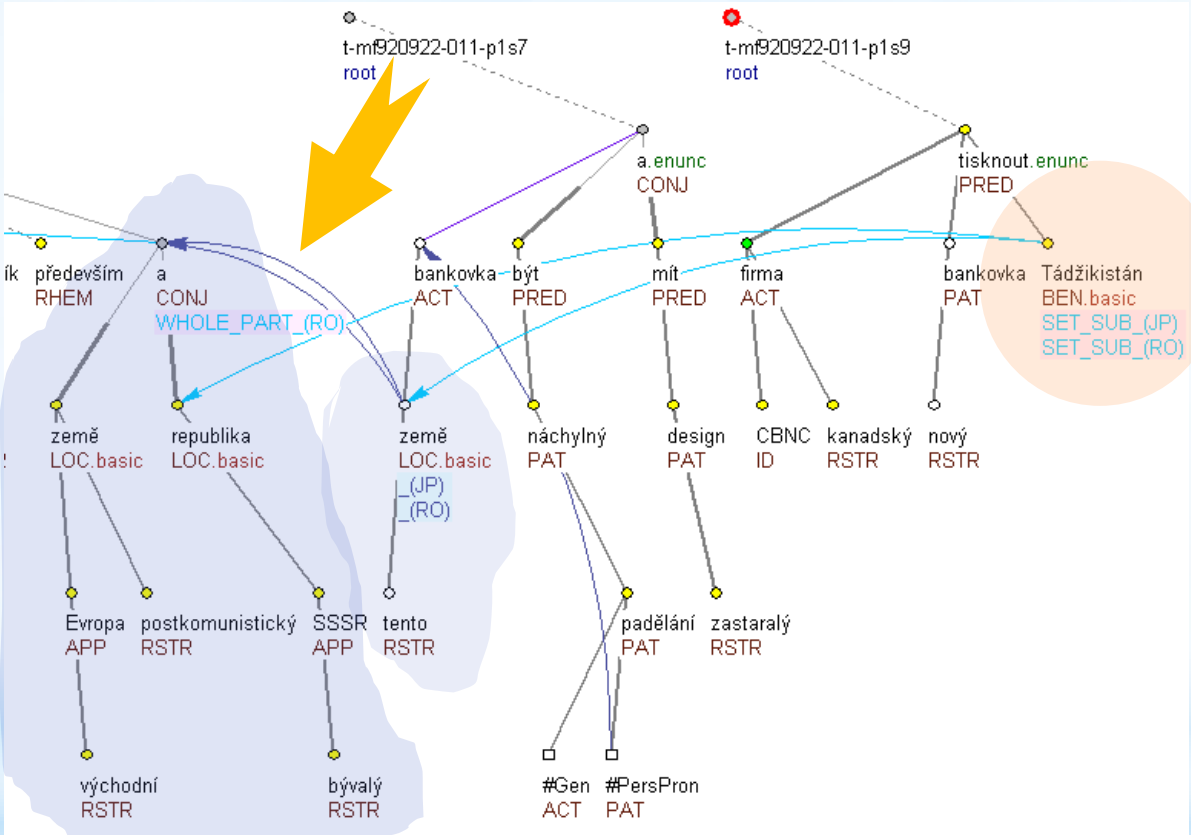
(=After the book had been already written, it was clear, that it is quite useful for parents too. The book contains explanations, how children go through divorce, how they react to it, and the instructions how parents should behave to minimize the suffering of their children..)

* Different selecting the antecedent/anaphoric element

Tiskárny bankovek mají i nové zákazníky, především v postkomunistických zemích východní Evropy a republikách bývalého SSSR. Bankovky v těchto zemích jsou náchylné na padělání a mají zastaralý design. Kanadská firma CBNC bude tisknout nové bankovky pro Tádžikistán

(= *They have new clients, first of all in the post-soviet countries of East Europe and in the republics of the former USSR. Banknotes in these countries can be easily falsified. The CBNC Company will print banknotes for Tajikistan.*)

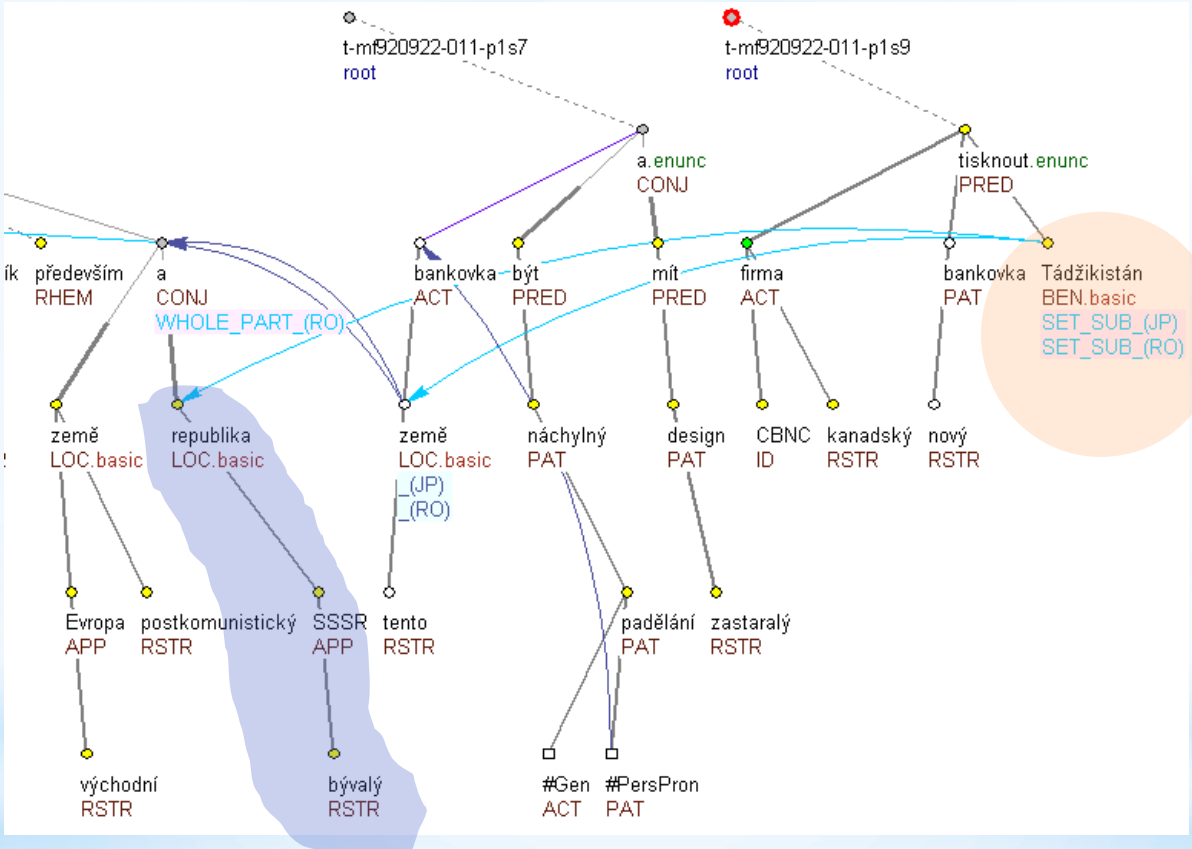
* Different selecting the antecedent/anaphoric element



Tiskárny bankovek mají i nové zákazníky, především v postkomunistických zemích východní Evropy a republikách bývalého SSSR. Bankovky v těchto zemích jsou náchylné na padělání a mají zastaralý design. Kanadská firma CBNC bude tisknout nové bankovky pro Tádžikistán .

(= *They have new clients, first of all in the post-soviet countries of East Europe and in the republics of the former USSR. Banknotes in these countries can be easily falsified. The CBNC Company will print banknotes for Tajikistan.*)

* Different selecting the antecedent/anaphoric element



Tiskárny bankovek mají i nové zákazníky, především v postkomunistických zemích východní Evropy a republikách bývalého SSSR. Bankovky v těchto zemích jsou náchylné na padělání a mají zastaralý design. Kanadská firma CBNC bude tisknout nové bankovky pro Tádžikistán

(= *They have new clients, first of all in the post-soviet countries of East Europe and in the republics of the former USSR. Banknotes in these countries can be easily falsified. The CBNC Company will print banknotes for Tajikistan.*)

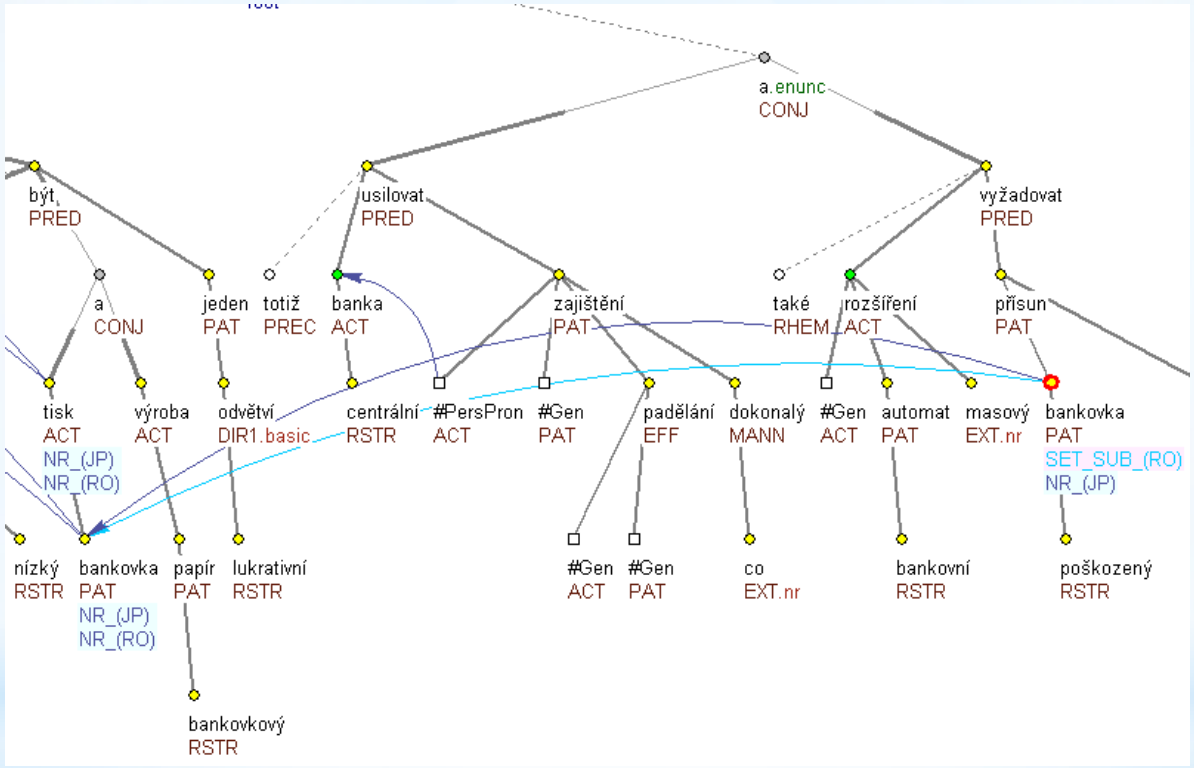
* Distinguishing between the bridging relations and the textual coreference

I přes klesající inflaci ve světě … je tisk bankovek a výroba bankovkového papíru jedním z nejlukrativnějších odvětví. […] *… Rozšíření bankovních automatů vyžaduje neustálý přísun nepoškozených bankovek.*

coreference (GEN) vs. bridging SUBSET

(= *Although inflation in the world rather decreases, … printing banknotes and production of banknote paper is still one of the most profitable areas. Mass expansion of ATMs calls for permanent increase of undamaged banknotes.*)

Distinguishing between the bridging relations and the textual coreference



I přes klesající inflaci ve světě … je tisk bankovek a výroba bankovkového papíru jedním z nejlukrativnějších odvětví. […] … Rozšíření bankovních automatů vyžaduje neustálý přísun nepoškozených bankovek.

(= Although inflation in the world rather decreases, … printing banknotes and production of banknote paper is still one of the most profitable areas. Mass expansion of ATMs calls for permanent increase of undamaged banknotes.)

*borderline cases between "specific" and "generic" coreference

U detergentu Toto jsme například řešili problém s udržením stálé kvality, protože jednotlivé partie byly nevyvážené. Investovali jsme dva miliony korun do nákupu pásových vah, zpřesnili dávkování a jakost pracího prášku stabilizovali.

engl. For example, for detergent Toto we thought about the problem of supporting the same quality .... We ... made the dosage more exact and so we set the quality of washing powder.

In ambiguous cases between specific and generic co-reference, we choose specific co-reference.

Začal jsem provozováním hospody, která byla mnohokrát vykradena. [... 2 věty ...] Hospoda byla jen startem, polem k podnikání s masem a masnými výrobky.

lit. engl. I began by carrying out a restaurant... [...] A/the restaurant was just the beginning [...]

*borderline cases between “specific” and “generic” coreference

K tématu pořadu TV NOVA TABU “Zrak za bílou hůl” byl přizván ke konzultaci Oldřich Čálek. Kateřina Hamrová, dramaturgyně pořadu, TV NOVA. (= *To consult the topic of the TV NOVA show TABU "Vision for a white cane", Oldřich Čálek was invited. Catherine Hamrová, the dramatist of the show, TV NOVA*)

Nic z toho se však nevyrovná míře neštěstí, které Romy postihlo v letech druhé světové války. Spolu se Židy byli označeni za méněcennou rasu a stali se objektem patologických fašistických opatření, jejichž cílem byla úplná genocida tohoto národa. (= *Nothing of this, however, compares to the misfortune that befell the Gipsies during the Second World War. Together with the Jews, they were called an inferior race and became the object of pathological fascist measures, their purpose being the complete genocide of the nation.*)

*Problem Cases - Reasons

different understanding of the content		
mostly don't have influence on understanding the text as a whole	"depth" of interpretation	guidelines "formalism"

Tak je knížka koncipována. V každé kapitole se mluví o určitém problému, uvádíme jak je rozsáhlý, kolik dětí je jím postiženo a co dělat. Je tam v podstatě konkrétní návod.
*This is the way **this book** is organised. **Every chapter** concerns a certain problem ... . There are actually specific instructions **there**.*

*I přes klesající inflaci ve světě … je tisk bankovek a výroba bankovkového papíru jedním z nejlukrativnějších odvětví. […] … Rozšíření bankovních automatů vyžaduje neustálý přísun nepoškozených bankovek. (= Although inflation in the world rather decreases, … printing **banknotes** and production of banknote paper is still one of the most profitable areas. Mass expansion of ATMs calls for permanent increase of **undamaged banknotes**.)*

Disagreement factors

- the text size
- degree of abstractedness of the text

Especially long texts with a large number of generic nouns, abstract and verbal nouns have the lowest inter-annotator agreement

- problematic are also
  - constructions with nouns of measure and time periods
  - generic noun phrases, abstract nouns and deverbatives
  - coreference between indefinite noun phrases

*Short text with 100% agreement

(1) ZLODĚJ SE VRÁTIL.

(2) Policejní hlídka vyrušila v neděli muže, který se vloupal do restaurace Kukačka v obci Horní Životice.

(3) Podařilo se mu zmizet, přestože policisté použili varovného výstřelu a vypustili služebního psa.

(4) Ještě téže noci se zloděj na místo činu vrátil.

(5) S policisty se tam Ø setkal podruhé.

(6) Tentokrát ho Ø zadrželi.

(7) Jedná se o několikrát trestaného M. K. z Ostravy.

Long text with low agreement

(11) Vaše kniha obsahuje ve třiadvaceti kapitolách různé problémy, od těžkých poškození dítěte až po lehčí disfunkci či vliv rozvodu na dítě.

(12) Tím ovšem jednu konkrétní rodinu může zajímat maximálně pět, přinejhorším deset kapitol.

(13) Zdeněk Matějček: Původně tato knížka byla určena pro zdravotnické pracovníky, a to především pro lékaře, kteří jsou ve styku s rodinou.

(14) Na druhé straně se ukázalo, že toto téma je stejně důležité pro pedagogy a vychovatele.

(15) Ti se přece setkávají i s postiženými nebo týranými dětmi.

(16) A když už byla knížka hotova, tak se zjistilo, že je praktická i pro rodiče.

(17) Samozřejmě ne každá kapitola ne pro každého rodiče.

(18) Zdeněk Dytrych: Kdyby se přímo dotýkalo některé rodiny deset kapitol, tak by to byla opravdu nešťastná rodina.

(19) Ale stačí jedna a většinou jich bude i víc.

(20) Vezměte si, kolik je rozvodů - třicet tisíc ročně v republice, to znamená, téměř třicet tisíc dětí je rozvodem nějakým způsobem postiženo.

(21) V této knize je poučení, jak snášejí děti rozvod a jak na něj reagují, a návod, jak se mají rodiče chovat, aby se utrpení dětí snížilo.

(22) Nebo například existuje lehká mozková disfunkce, kterou trpí podle našeho rozsáhlého výzkumu pět procent dětí.

(23) Toto postižení se velice špatně rozpoznává.

(24) Dítě je nemotorné, neklidné a není schopné se soustředit, ale přitom je většinou chytré.

(25) Rodiče ho považují za lajdáka a bývá trestáno třeba za špatný výkon ve škole, tím se zhoršuje vztah k učení atd.

(26) A tohle rodiče musí vědět.

(27) Samozřejmě i pedagogové a v této knížce je návod co s tím.

(28) Zdeněk Matějček: Předkládáme i problémy, na které se zapomíná.

(29) Tak například úmrtí dítěte nebo narození postiženého dítěte.

(30) Tady nejde jenom o rodiče, ale i o okolí, které musí vědět, jak se má chovat.

(31) Nebo úmrtí v rodině a jeho vliv na dítě a může to být třeba babička.

*Reasons for Disagreement – Abstract Nouns

one of very weak points in the PDT coreference annotaion

- attempted in PDT, also classified for specific and generic abstracts (e.g. according to the reference of valencies)
- actually my problem was that I couldn't reliably separate abstract nouns from concrete ones

*Preferuji širší předvedení s mnoha vnitřními souvislostmi, protože nám chybějí kritéria pro hodnocení **současné české výtvarné kultury**.* {… 11 sentences inbetween…} *Měli bychom se znovu pokusit … získávat **současné umění**, abychom jednou měli autentický soubor naší doby* (= *I prefer wider demonstration with many internal connections because we lack criteria for evaluation of contemporary Czech art*. {… 11 sentences inbetween …} *We should try ... to acquire the contemporary art again, in order to get an authentic set of our time.*)

antecedent is relatively far from the anaphoric NP

Reasons for Disagreement – Abstract Nouns

*Tímto faktorem je podnikatel-inovátor, který se snaží o **zisk**, a proto logicky nemůže existovat ve stavu statiky, která nezná ani **zisk**, ani ztrátu.* (= *This factor is the enterpreneur-innovator, who is trying to gain profit, and hence, logically, cannot exist in a static state, where there is no profit or loss.*)

*Ve specifických podmínkách **české ekonomiky** růst nezaměstnanosti v letech 1991-1993 značně zaostal za poklesem HDP. […] Nejméně dvouprocentní růst **české ekonomiky** již letos.* (=*In the specific conditions of the Czech economy the growth of unemployment... This year at least a two percent growth of the Czech economy.*)

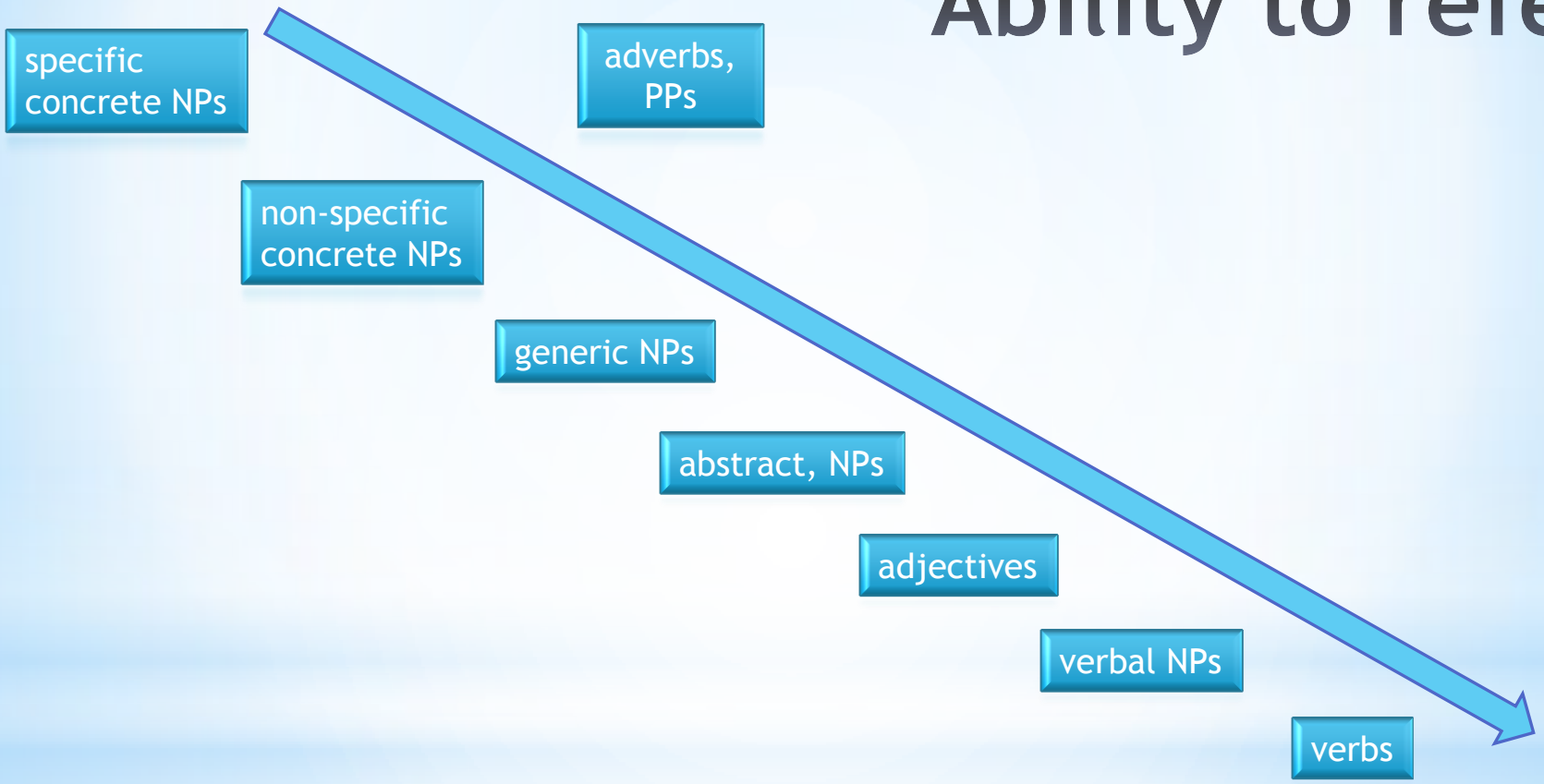
*In the Treasury market, investors paid scant attention to the day's economic reports, which for the most part provided a mixed view of **the economy**. ``Whether you thought **the economy** was growing weak or holding steady, yesterday's economic indicators didn't change your opinion," said Charles Lieberman, a managing director at Manufacturers Hanover Securities Corp.*

Reasons for Disagreement – Verbal Nouns

*Vedení Pojišťovny Investiční a Poštovní banky nás upozornilo, že jejich pojišťovna nebyla zařazena mezi ty, které umožňují úrazové připojištění, ač tuto službu poskytují. Omlouváme se za **toto nedopatření**, dotyčná redaktorka byla pokutována.* (=*The Insurance Investment and the Post Bank management has notified us that their insurance company was not included among those that allow casualty insurance, although it provides this service. We apologize for this oversight, the editor who made the mistake was fined.*)

*Rychlé, avšak i bezpečné **vypořádání**. Rychlost **vypořádání** burzovních obchodů v čase odpovídá podle Jiřího Béra potřebám.* (= *Fast, yet safe transaction. According to Jiřího Bér's opinion, the speed of transaction corresponds to the needs.*)

*Ability to refer



* Reasons for Disagreement – measure NPs and other NPs with a ‘container’ meaning

skupina lidí (= a group of people)

počet akcií (= a number of stocks)

stádo krav (= a herd of cows)

dostatek financí (= abundance of finances)

milióny Židů (= millions of Jews)

sklenice piva (= a glass of beer)

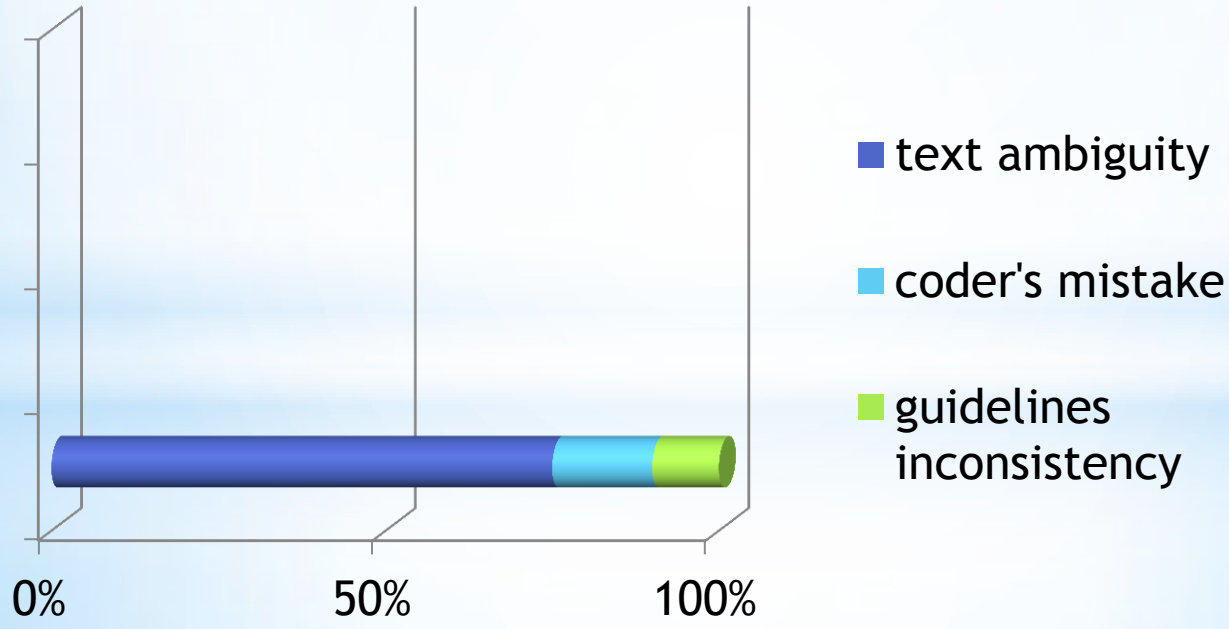
deset procent obyvatel (= ten percent of population)

* Reasons for Disagreement – Constructions with Time Periods

That compares with operating earnings of $132.9 million, or 49 cents a share, ***the year earlier****.*
The prior-year period *includes...*

* Reasons for Disagreement

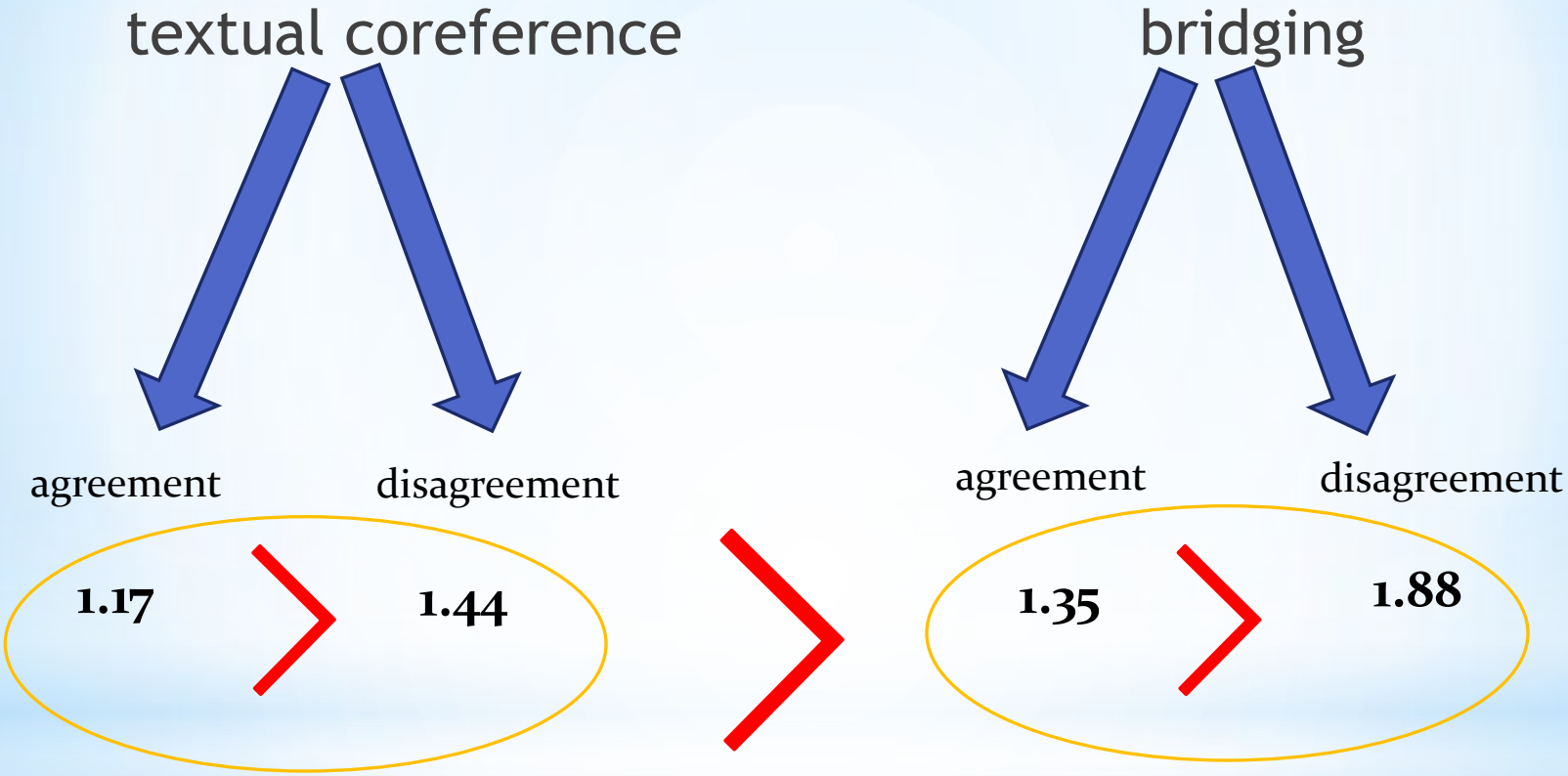
almost three fourth of the coders' disagreements come from the text ambiguity (empirically ambiguous or near-identical in the sense of Recasens (2010))



Experiment - Certainty of the manual annotations

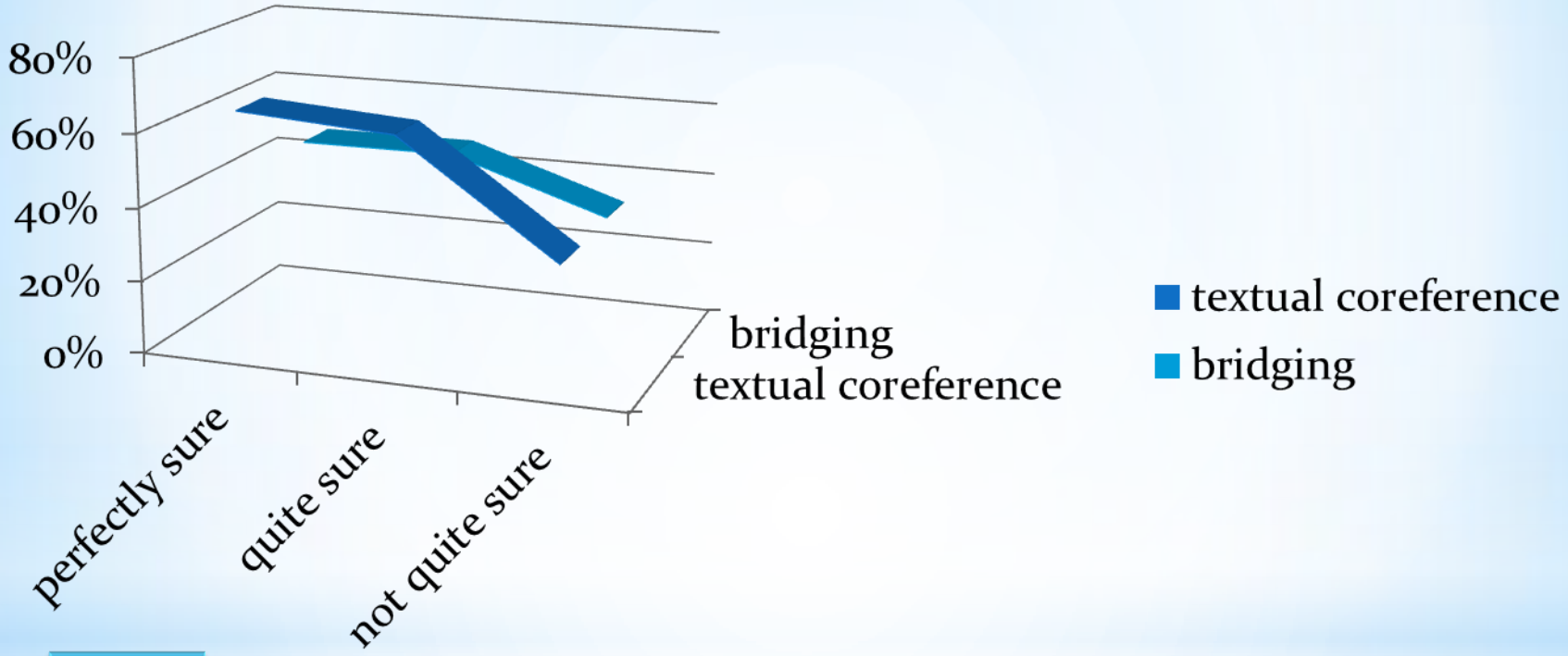
- annotators marked the certainty for their annotation decisions on the scale of 1 to 3
  - 1 : perfectly sure,
  - 2 : quite sure,
  - 3 : not quite sure
- certainty marked for
  - the presence of a relation,
  - selecting the antecedent,
  - distinguishing between the bridging relation and the textual coreference and
  - selecting the type of the bridging relation or the textual coreference

*Certainty in the Presence of a Relation



- naturally, the lower the agreement is, the less are the annotators sure
- the number of cases where the annotators didn't mark uncertainty but still disagreed exceeds all other cases (56 disagreements, only 26 were marked)

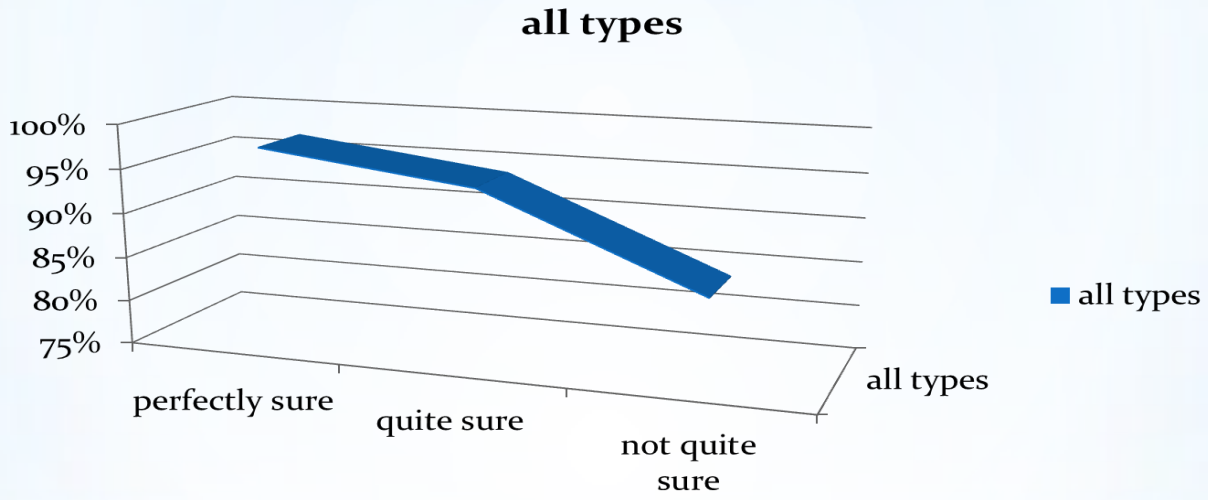
*Certainty in selecting the Antecedent



Again:

- the numbers show a lower agreement in cases where the annotators were not sure about the antecedent BUT
- from 27 disagreements in choosing the antecedent, only 16 were marked as uncertain by at least one annotator

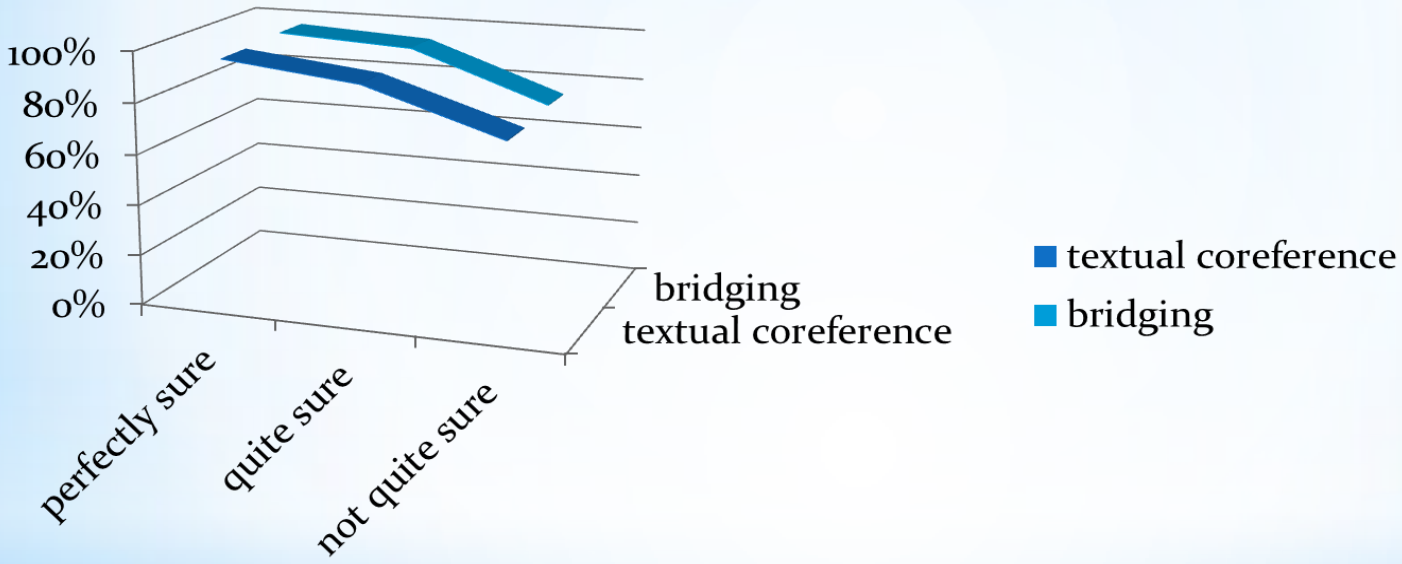
*Certainty in Distinguishing Between Coreference and Bridging



Numbers show:

- The difference in agreement between “certain” and “uncertain” relations in this case is not so relevant
- In most cases (21 of 32), the annotators marked ambiguity but still made the same decision

*Certainty in selecting the type of bridging or coreference relation



Again:

the numbers show a lower agreement in cases where the annotators were not sure about the type of the relation

* Results of the experiment and analysis

inter-annotator agreement + annotators' certainty reveal:

* empirical ambiguity is much more frequent on text level than on syntax level and lower
* the complexity of real corpus data which can never be reflected by any annotation guidelines
* ambiguity is frequently not detected by annotators
* in many cases world knowledge is needed
* annotators are more sure about relations between noun phrases in topic and contrastive topic than about those in focus

*Future plans?

Future plans

* Coreference chains - analyzing coreference chains as concerns realization of referring expressions in texts, number of chains in texts of different length and genres, length of chains and so on
* Coreference chains in different languages CS – EN – other languages (Russian, Polish, German)
* Predictability of realization of anaphoric expressions
* Linguistic analysis of coreferential counterparts (pro-drop qualities, possessivity, finite vs. non-finite, correlative) CS – EN – other languages (Russian, Polish, German)
* Working with parallel data (translated vs. original texts)
* Coreference projection CS – EN – other languages (Russian, Polish, German)
* Discourse annotation projection (CS - EN)
* Multi-language comparison of anaphoric and discourse relations
* Comparative theoretical and NLP + corpus analysis of anaphoric connectives
* Direction pragmatics (exophoric pronouns, demonstrative pronoun use)
* Saliences (beeing done) – analyzing the degree of activation and realization of referents in texts

Pilot study - Coreference chains in English, Czech and Russian

	PCEDT	RuCor
economics texts	161	166
political news	230	231
other news	112	105
TOTAL	503	499

Chain length	English	Czech	Russian
TOTAL number of chains	420 (100.0%)	485 (100.0%)	263 (100.0%)
number of 2-elements chains	254 (60.5%)	304 (62.7%)	139 (52.9%)
number of chains of length 3-4	108 (25.7%)	109 (22.5%)	64 (24.3%)
number of chains of length 5-8	39 (9.3%)	47 (9.7%)	33 (12.5%)
number of chains longer than 8	19 (4.5%)	25 (5.1%)	28 (10.6%)

Pilot study - Coreference chains in English, Czech and Russian

NP type \ language		English		Czech		Russian	
central pronouns	subj	121	8.6%	2	0.1%	39	3.8%
	non-subj	129	9.2%	125	7.8%	95	9.3%
relative		96	6.9%	136	8.5%	42	4.1%
anaphoric zero		28	2.0%	304	19.1%	13	1.3%
bare noun		75	5.4%	208	13.1%	164	16.0%
NP with determiner		315	22.5%	63	4.0%	20	1.9%
NP with other modif		119	8.5%	313	19.7%	172	16.8%
NP including NE		104	7.4%	141	8.9%	80	7.8%
NE		331	23.7%	216	13.6%	379	37.0%
other		81	5.8%	83	5.2%	21	2.0%
TOTAL		1399	100%	1591	100%	1025	100%

*Acknowledgements

* The research was supported from the Grant Agency of the Czech Republic (grant P406/12/0658 Coreference, discourse relations and information structure in a contrastive perspective). This work has been using language resources developed, stored, and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).
* I gratefully acknowledge my colleagues Eva Hajičová, Šárka Zikánová, Lucie Poláková, Jiří Mírovský, Kateřina Rysová et al.