Discourse and Coherence: A Complex Approach to Textual Phenomena in the Prague Dependency Treebank

Анна Недолужко
Карлов университет, Прага

Связность текста: Комплексный подход к дискурсивным явлениям в Пражском синтаксически размеченном корпусе

Анна Недолужко
Карлов университет, Прага

План

I. Корпуса и дискурсивные явления
II. Пражский корпус – разметка
III. Исследования – прикладные, статистико-дескриптивные и лингвистические. Примеры.
    a. Пример 1: имплицитные анафорические отношения
    b. Пример 2: имплицитные дискурсивные отношения
IV. Выводы

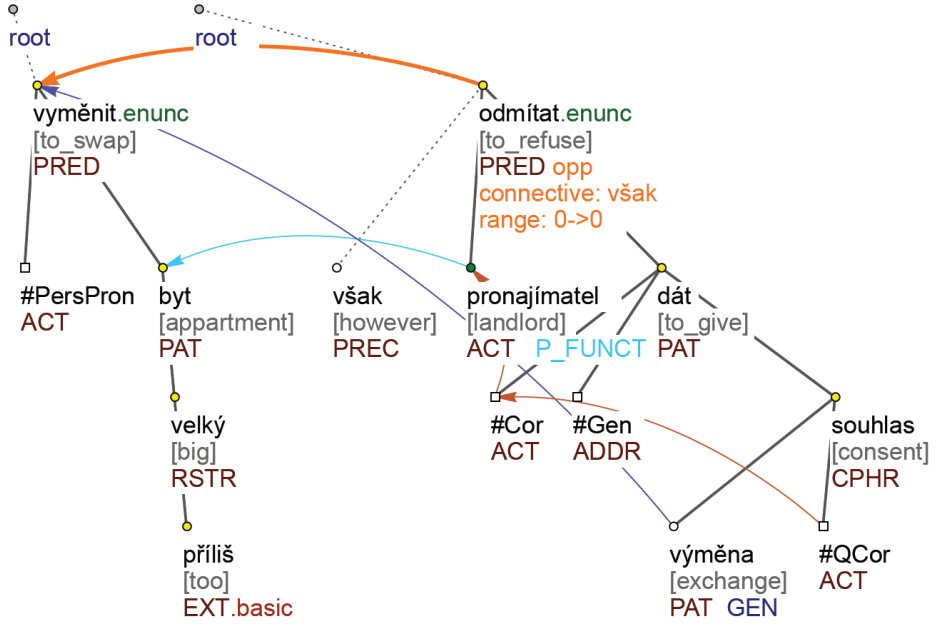
Figure from Taboada (2015)

Текст

Текстура

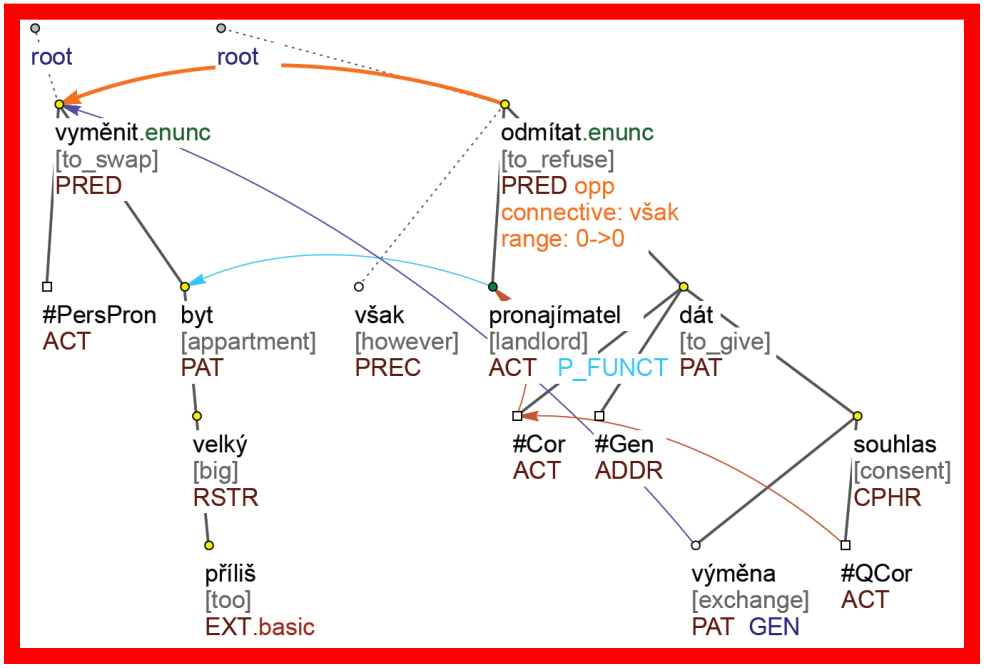
Halliday and Hasan (1976)

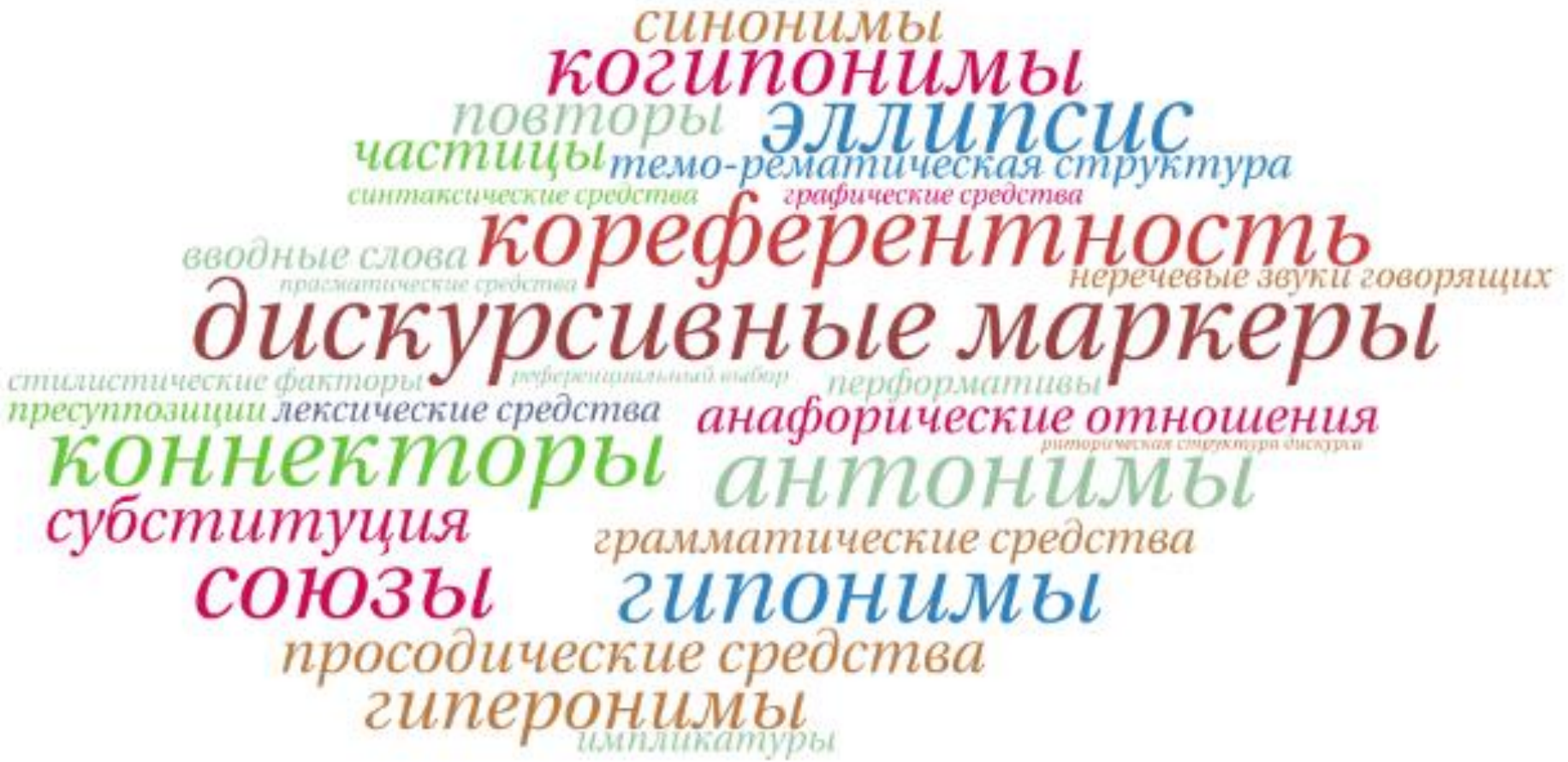
Языковые ресурсы



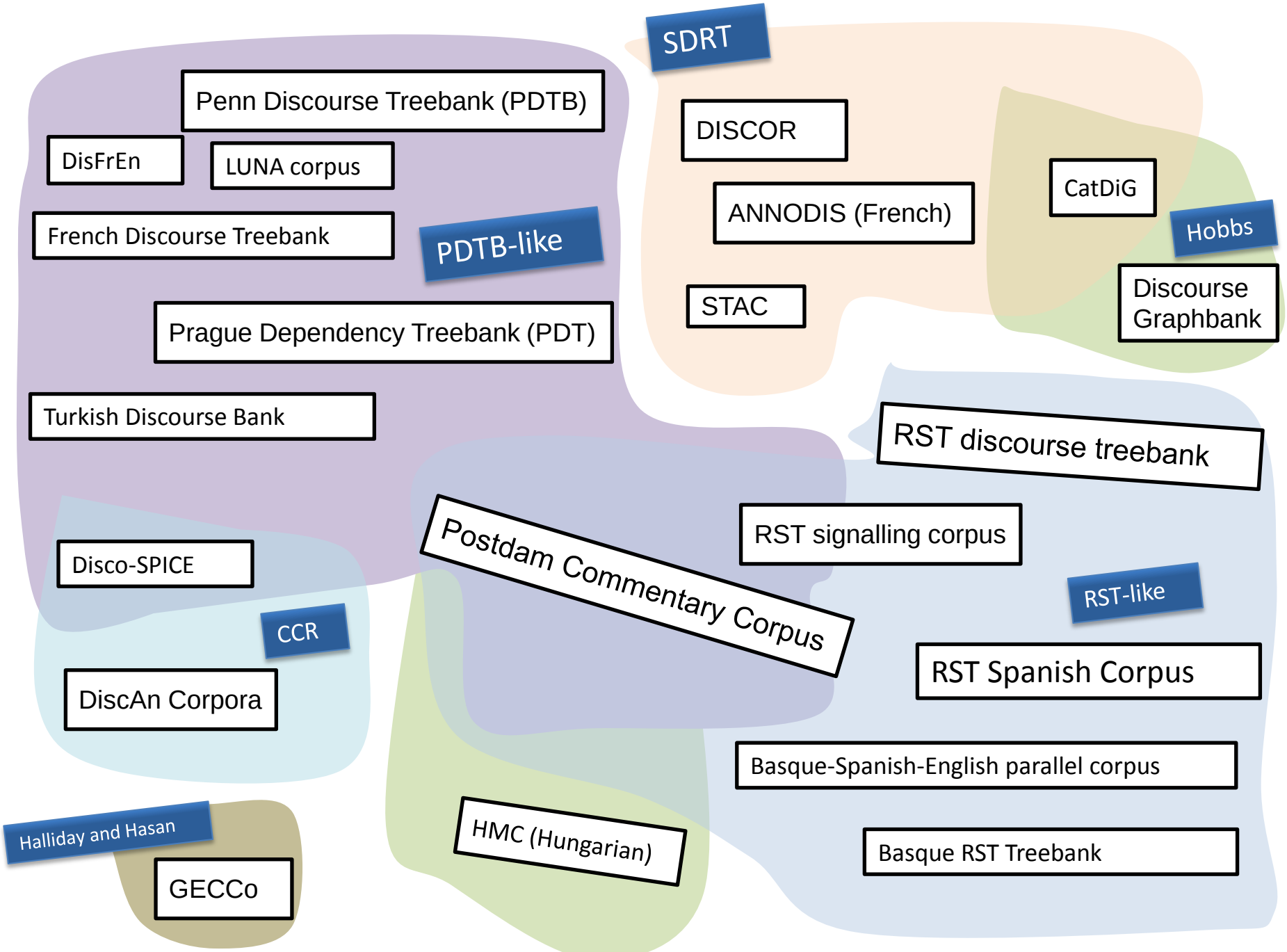
Языковые ресурсы

- sampling
- парадигма (многозначность)
- сомнительность сложных данных




синонимы
когипонимы
повторы
эллипсис
частицы
темо-рематическая структура
синтаксические средства
графические средства
кореферентность
вводные слова
неречевые звуки говорящих
прагматические средства
дискурсивные маркеры
стилистические факторы
референциальный выбор
перформативы
пресуппозиции
лексические средства
анафорические отношения
риторическая структура дискурса
коннекторы
антонимы
субституция
грамматические средства
союзы
гипонимы
просодические средства
гиперонимы
импликатуры

- Дискурсивная разметка (коннекторы, союзы, структура дискурса)
- Кореферентность и анафорические отношения


SDRT
Penn Discourse Treebank (PDTB)
DISCOR
DisFrEn
LUNA corpus
CatDiG
ANNODIS (French)
Hobbs
French Discourse Treebank
PDTB-like
Discourse Graphbank
STAC
Prague Dependency Treebank (PDT)
Turkish Discourse Bank
RST discourse treebank
Postdam Commentary Corpus
RST signalling corpus
Disco-SPICE
RST-like
CCR
RST Spanish Corpus
DiscAn Corpora
Basque-Spanish-English parallel corpus
Halliday and Hasan
HMC (Hungarian)
Basque RST Treebank
GECCo

Разметка дискурса

- Penn Discourse Treebank (Prasad et al.) +EntRels
- RST Discourse Treebank (Carlson et al. 2003)
- CatDiG (Badia et al.)
- RST signalling corpus (Dag&Taboada)
- Postdam Commentary Corpus (Neumann, Stede)
- DISCOR and ANNODIS (Afantenos et al.)
- PragueDT (discourse: Zikánová, Poláková et al.)
- DiscAn (Sanders et al.)
- GECCo (Kunz et al.) ...

Разметка кореферентности

- MUC corpora (Hirschman, 1998)
- ACE corpora (Doddington et al. 2000)
- Ontonotes (Pradhan et al.)
- AnCora (Recasens et al.)
- MATE, (Poesio et al.)
- GNOME (Poesio et al.)
- ARRAU (Poesio et al.)
- Postdam Commentary Corpus (coref. Stede, Grishina et al.)
- Prague Dependency Treebank (coref. Nedoluzhko et al.)
- RuCor (Toldova et al.)
- Polish Coreference corpus (Ogrodniczuk et al.)
- GECCo (Kunz et al.)

Разметка дискурса

- Penn Discourse Treebank (Prasad et al.) +EntRels
- RST Discourse Treebank (Carlson et al. 2003)
- CatDiG (Badia et al.)
- RST signalling corpus (Dag&Taboada)
- Postdam Commentary Corpus (Neumann, Stede)
- DISCOR and ANNODIS (Afantenos et al.)
- PragueDT (discourse: Zikánová, Poláková et al.)
- DiscAn (Sanders et al.)
- GECCo (Kunz et al.)

...

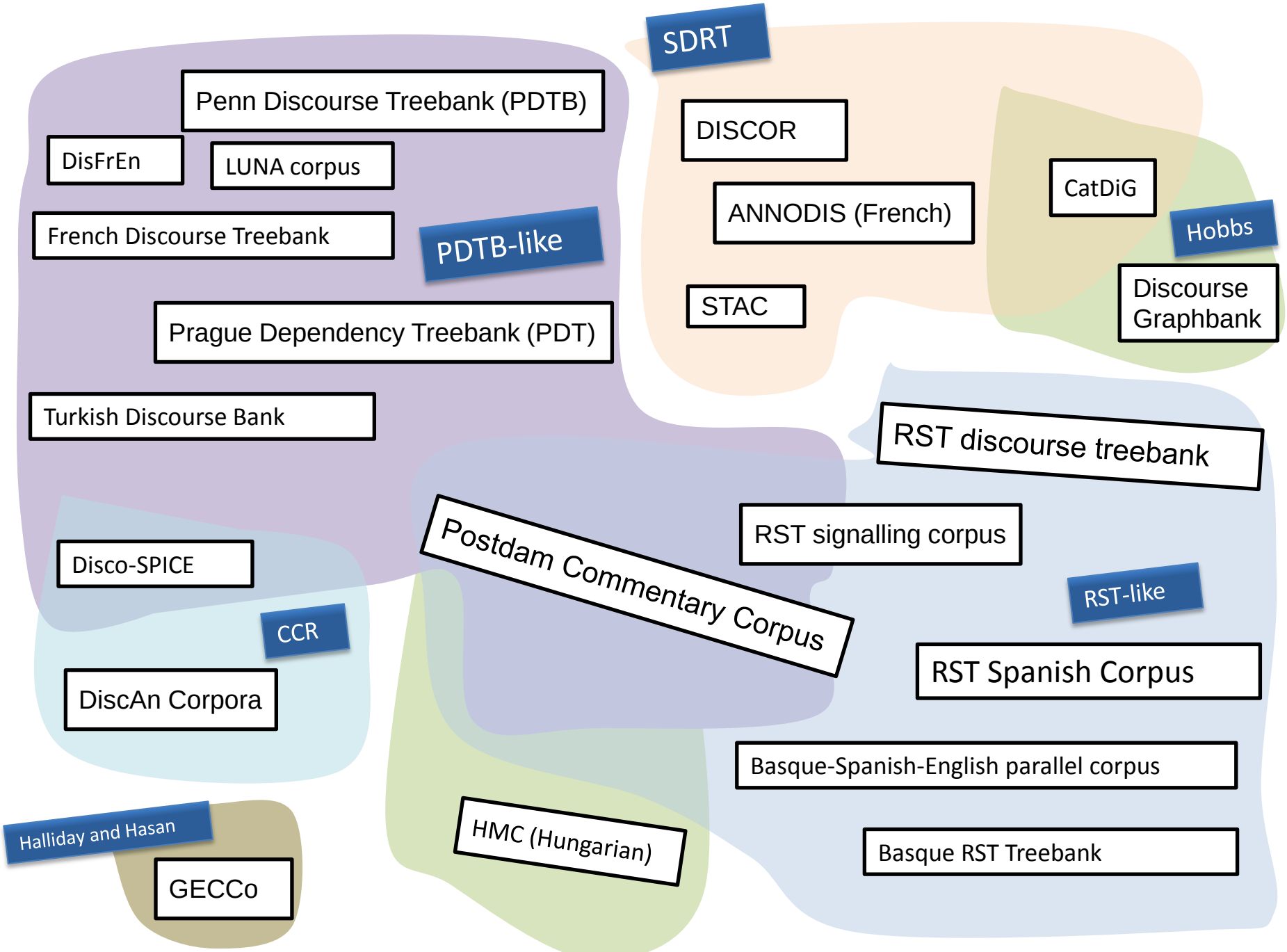
Разметка кореферентности

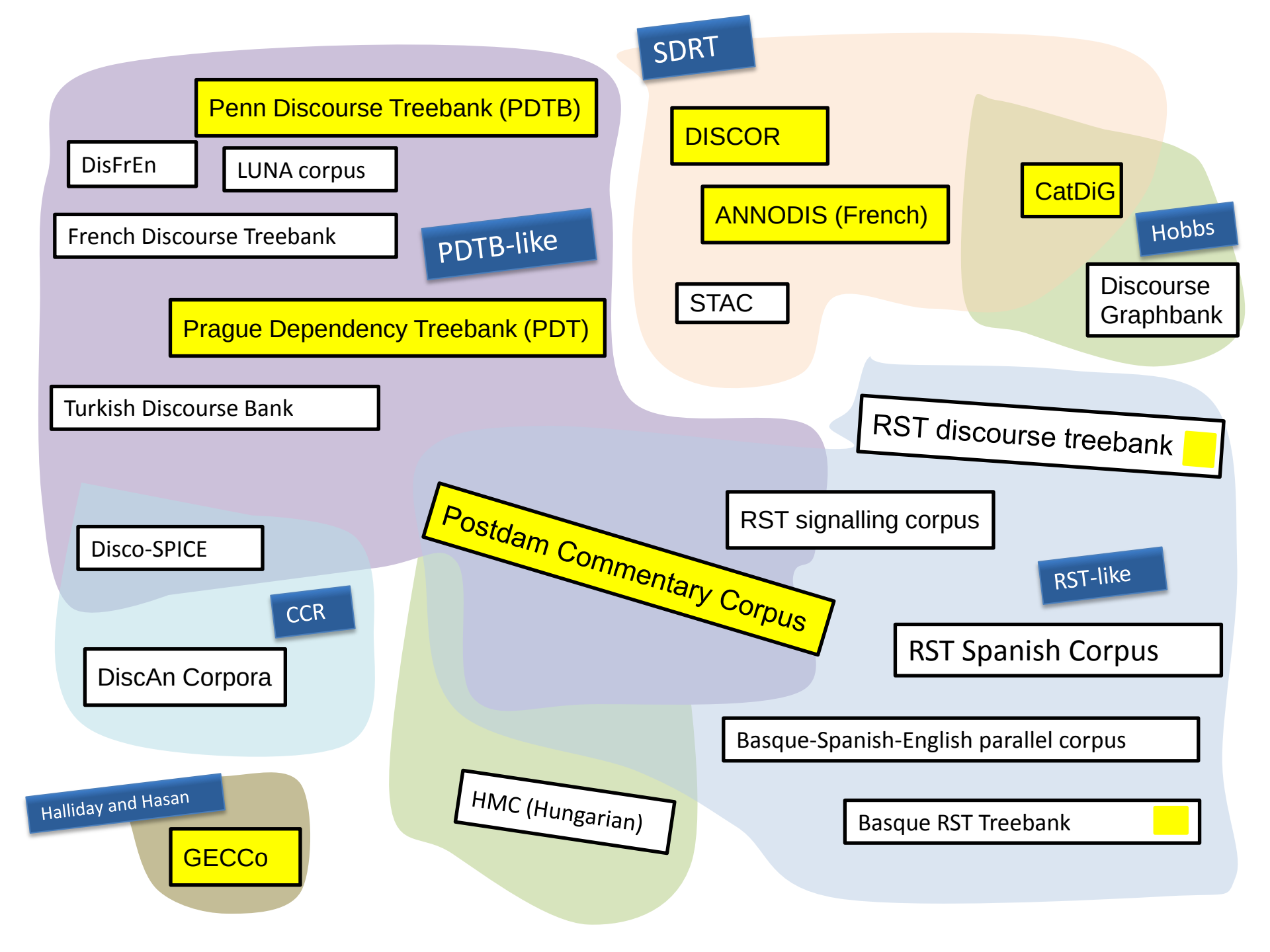
- MUC corpora (Hirschman, 1998)
- ACE corpora (Doddington et al. 2000)
- Ontonotes (Pradhan et al.)
- AnCora (Recasens et al.)
- MATE, (Poesio et al.)
- GNOME (Poesio et al.)
- ARRAU (Poesio et al.)
- Postdam Commentary Corpus (coref. Stede, Grishina et al.)
- Prague Dependency Treebank (coref. Nedoluzhko et al.)
- RuCor (Toldova et al.)
- Polish Coreference corpus (Ogrodniczuk et al.)
- GECCo (Kunz et al.)

Разметка дискурса

- Penn Discourse Treebank (Prasad et al.) +EntRels
- RST Discourse Treebank (Carlson et al. 2003)
- CatDiG (Badia et al.)
- RST signalling corpus (Dag&Taboada)
- Postdam Commentary Corpus (Neumann, Stede)
- DISCOR and ANNODIS (Afantenos et al.)
- PragueDT (discourse: Zikánová, Poláková et al.)
- DiscAn (Sanders et al.)
- GECCo (Kunz et al.)

...

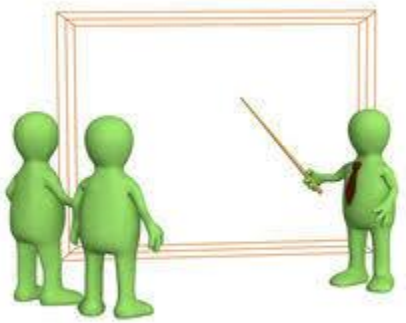

SDRT
Penn Discourse Treebank (PDTB)
DISCOR
DisFrEn
LUNA corpus
CatDiG
ANNODIS (French)
Hobbs
French Discourse Treebank
PDTB-like
Discourse Graphbank
STAC
Prague Dependency Treebank (PDT)
Turkish Discourse Bank
RST discourse treebank
Postdam Commentary Corpus
RST signalling corpus
Disco-SPICE
RST-like
CCR
RST Spanish Corpus
DiscAn Corpora
Basque-Spanish-English parallel corpus
Halliday and Hasan
HMC (Hungarian)
Basque RST Treebank
GECCo


SDRT
Penn Discourse Treebank (PDTB)
DISCOR
DisFrEn
LUNA corpus
CatDiG
ANNODIS (French)
Hobbs
French Discourse Treebank
PDTB-like
Discourse Graphbank
STAC
Prague Dependency Treebank (PDT)
Turkish Discourse Bank
RST discourse treebank
Postdam Commentary Corpus
RST signalling corpus
Disco-SPICE
RST-like
CCR
RST Spanish Corpus
DiscAn Corpora
Basque-Spanish-English parallel corpus
Halliday and Hasan
HMC (Hungarian)
Basque RST Treebank
GECCo

Prague Dependency Treebank (PDT 3.0)

- чешские газетные тексты
- многоуровневая лингвистическая разметка
- 3165 текстов (= 49431 предложение = 833195 токенов)

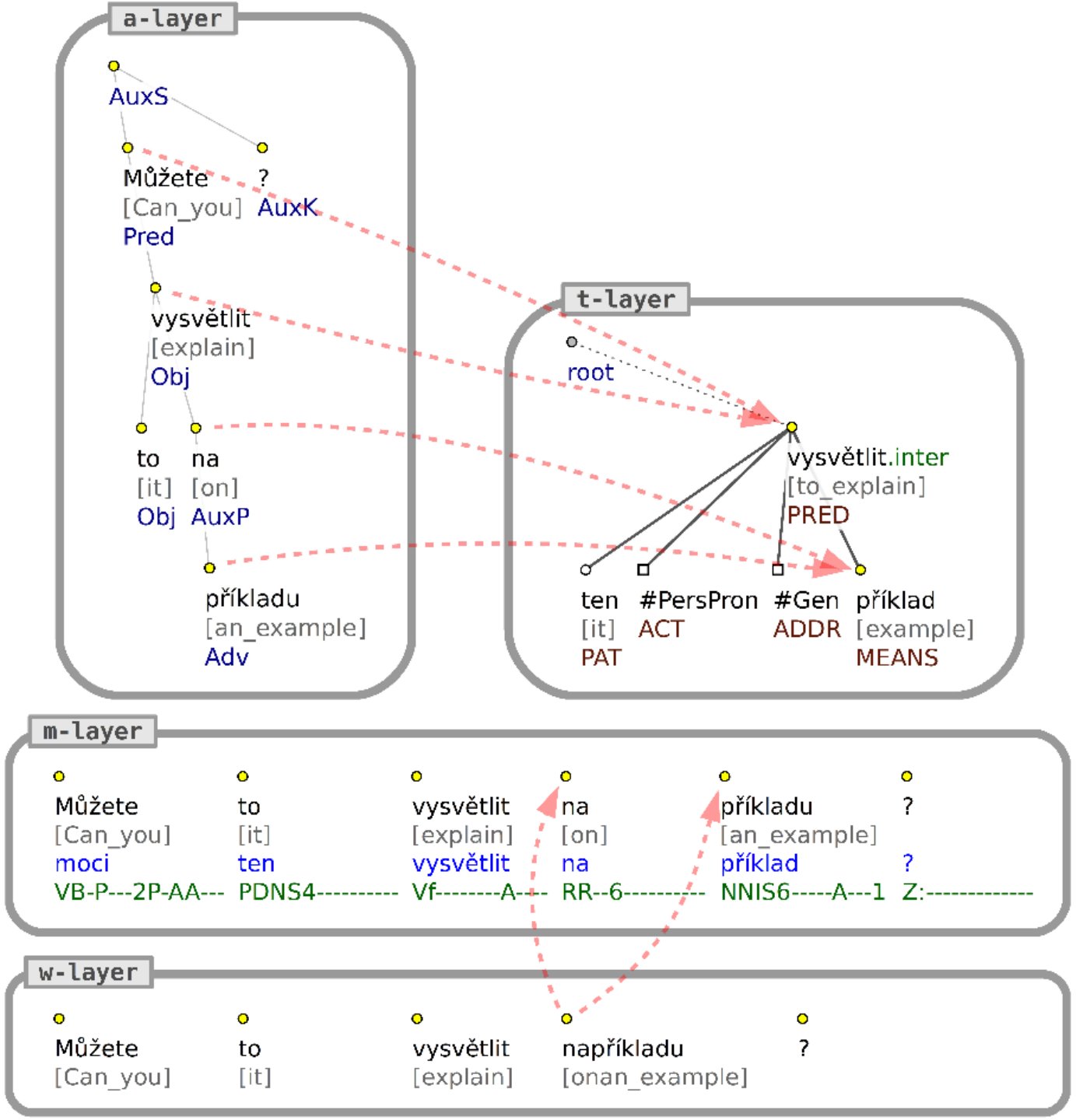
Пример:



чеш. ***Můžete to vysvětlit na příkladu?***

рус. *(Вы) можете это объяснить на (каком-нибудь) примере?*

англ. *Can you explain it on an example?*


a-layer
AuxS
Můžete
[Can_you]
Pred
?
AuxK
vysvětlit
[explain]
Obj
to
[it]
Obj
na
[on]
AuxP
příkladu
[an_example]
Adv
t-layer
root
vysvětlit.inter
[to_explain]
PRED
ten
[it]
PAT
#PersPron
ACT
#Gen
ADDR
příklad
[example]
MEANS
m-layer
Můžete
[Can_you]
moci
VB-P---2P-AA---
to
[it]
ten
PDNS4----------
vysvětlit
[explain]
vysvětlit
Vf-------A---
na
[on]
na
RR--6----------
příkladu
[an_example]
příklad
NNIS6-----A---1
?
?
Z:----------
w-layer
Můžete
[Can_you]
to
[it]
vysvětlit
[explain]
napříkladu
[onan_example]
?

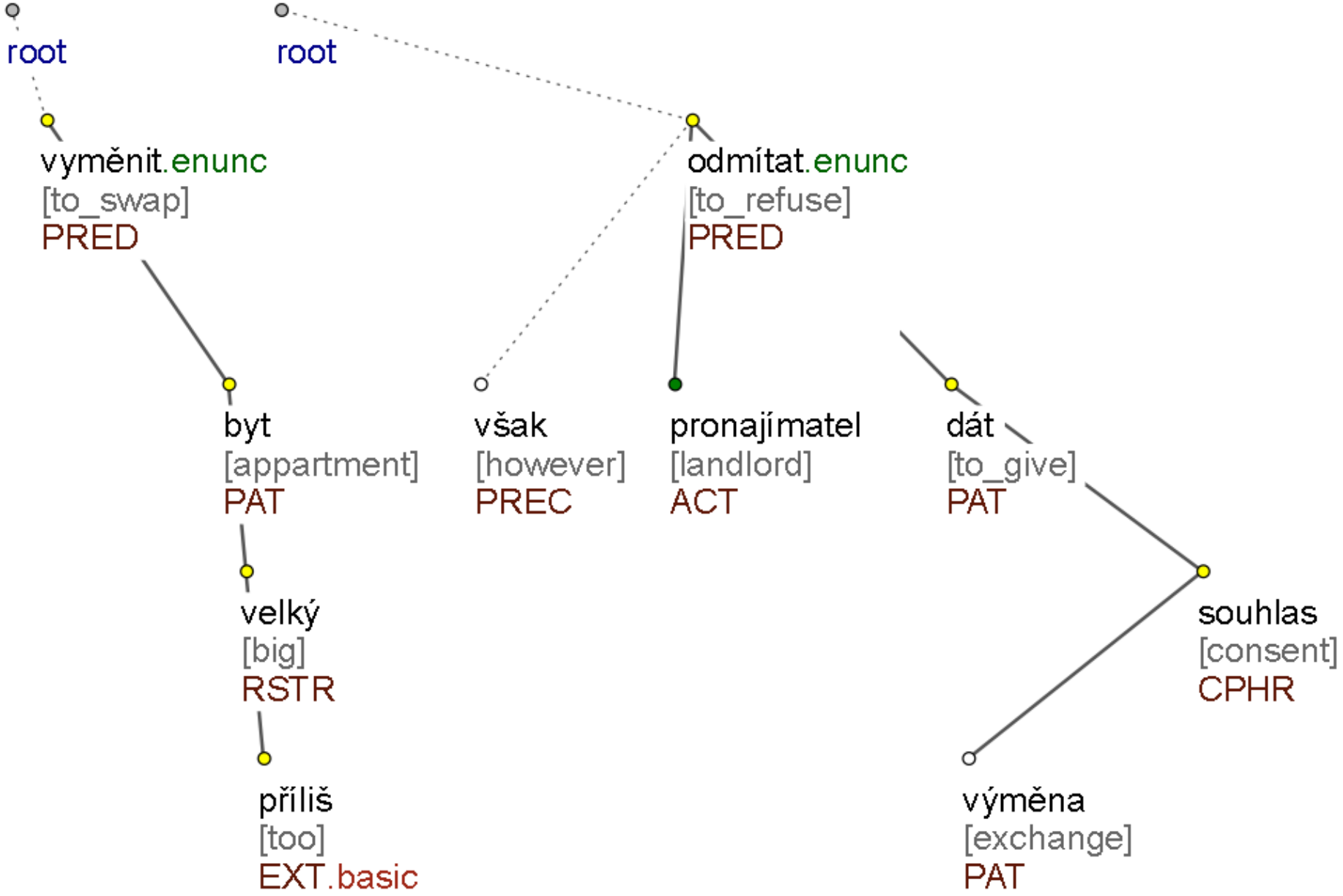
Разметка на уровне текста

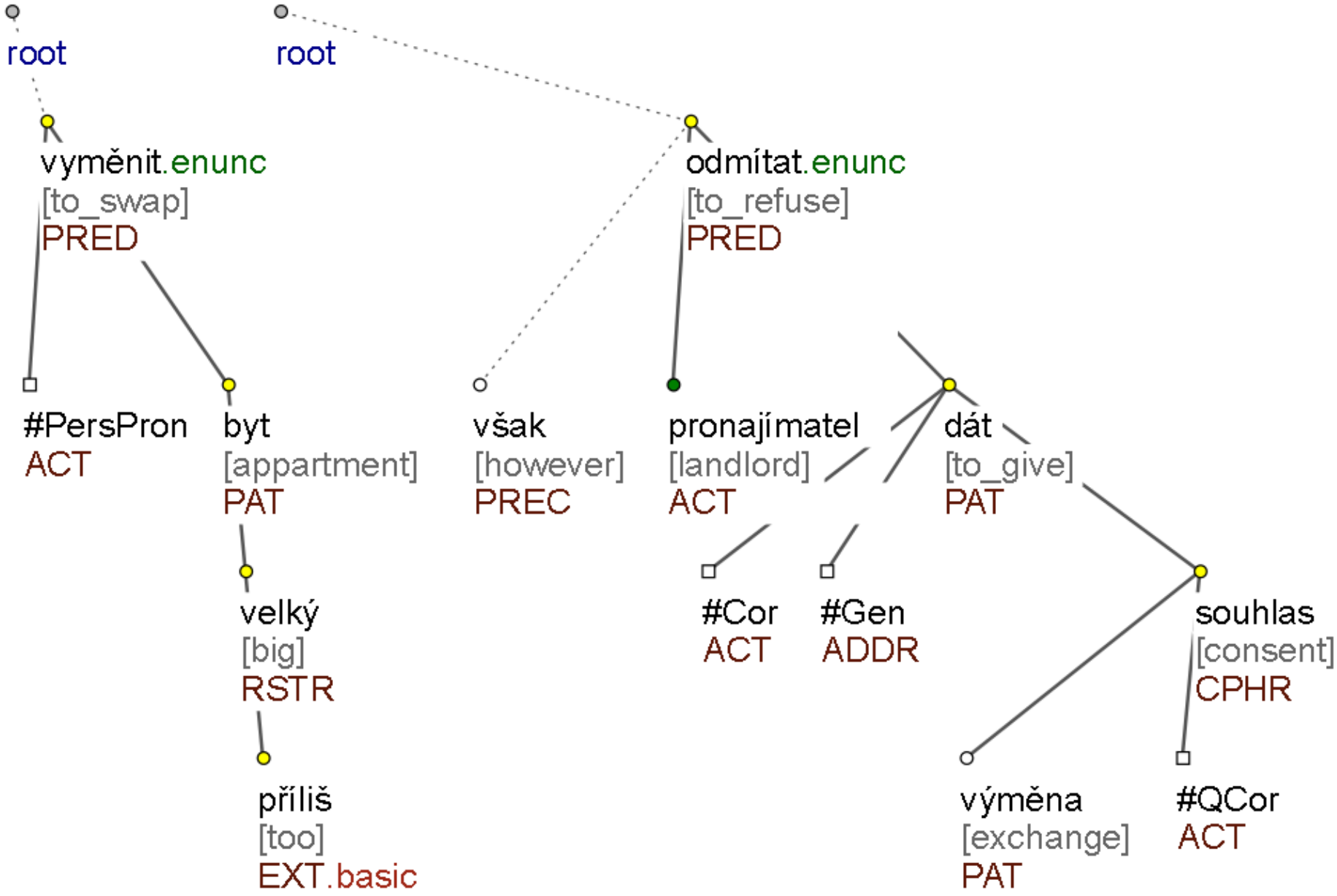
- **кореферентность**
- **ассоциативная анафора** (бриджинг)
- **актуальное членение** (tfa)
- **контекстная учтенность**, активированность в памяти говорящего
- **дискурсивная** разметка: дискурсивные маркеры (коннекторы), аргументы связанные этими коннекторами и типы отношений между ними

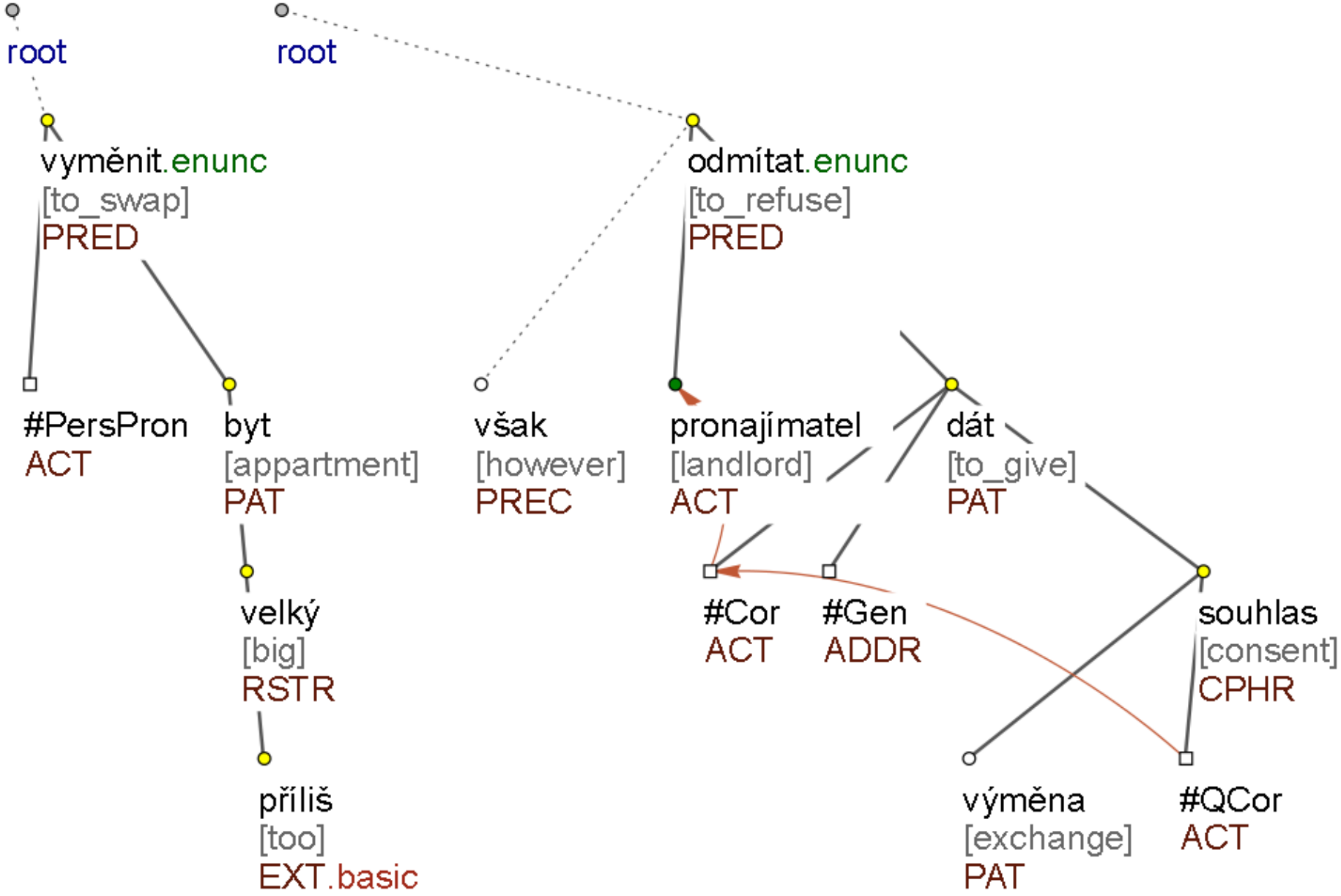
Все явления размечены отдельно различными аннотаторами!!!

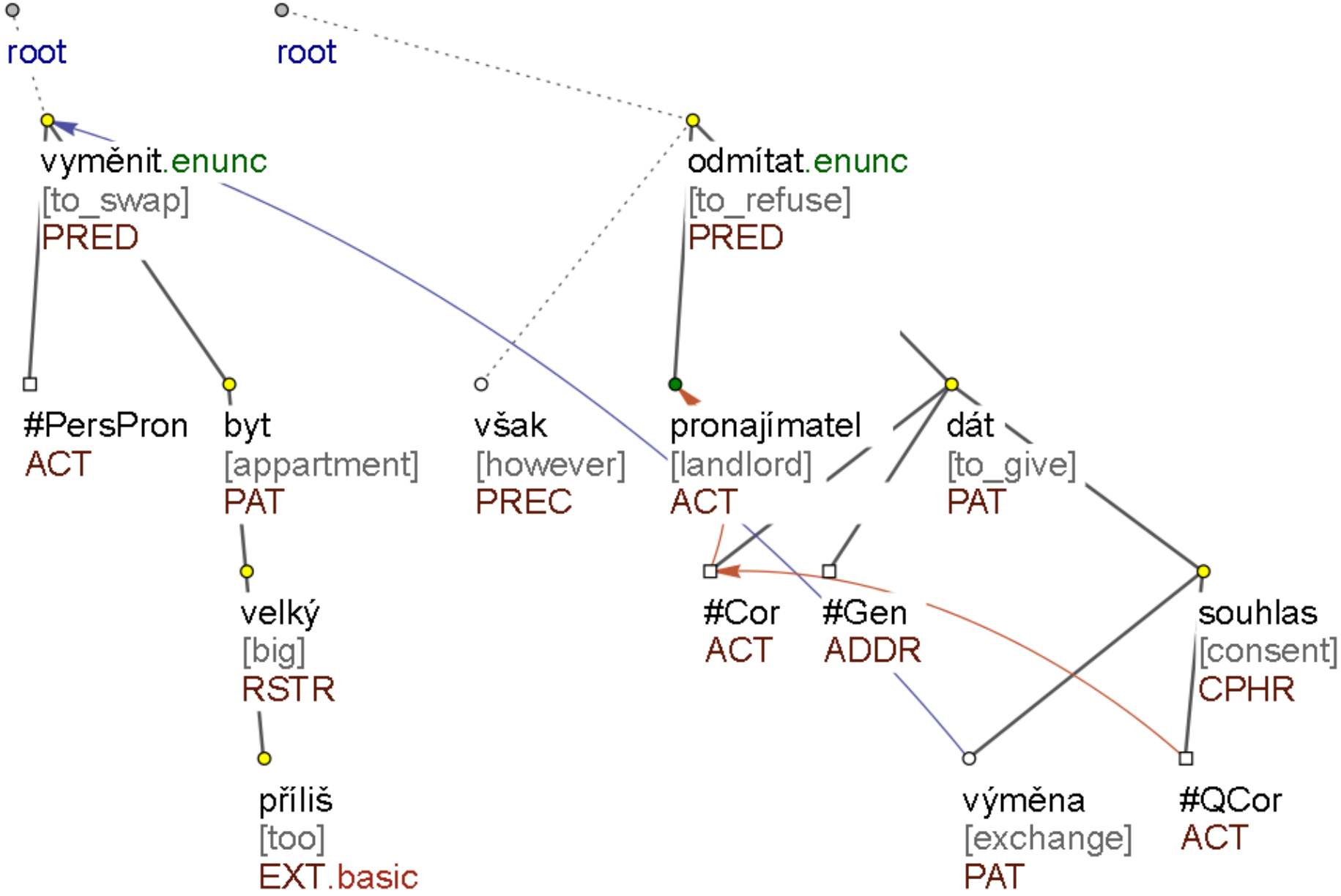
Пример:

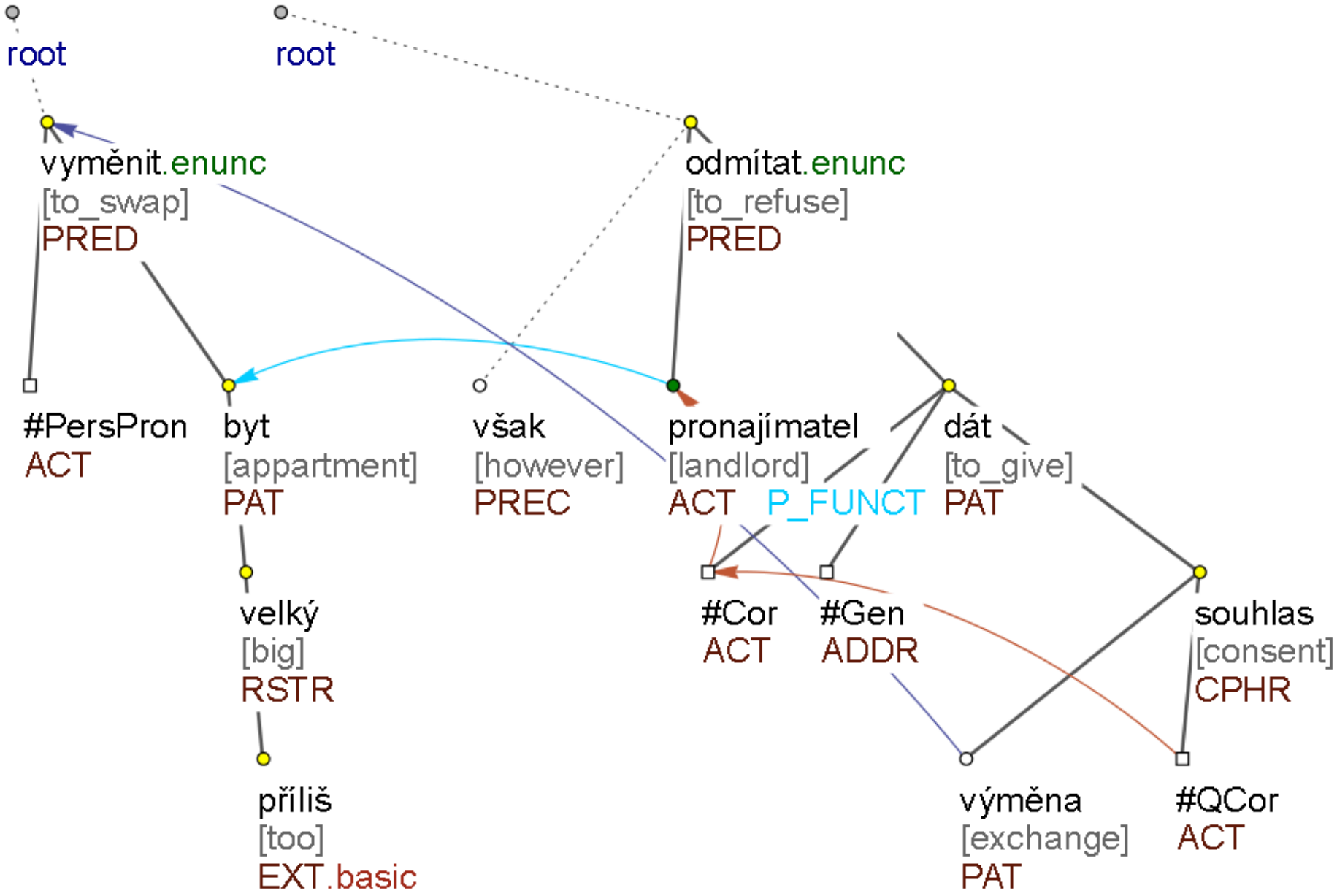
- чеш. *Chtěl bych vyměnit příliš velký byt. Pronajímatel však odmítá dát k výměně souhlas*.
- рус.(досл.): (*Я) хотел бы обменять слишком большую квартиру*. *Хозяин (квартиры) однако отказывается дать согласие на обмен*.
- англ. *I would like to swap an apartment that is too big. But the landlord refuses to give his consent for the exchange*.

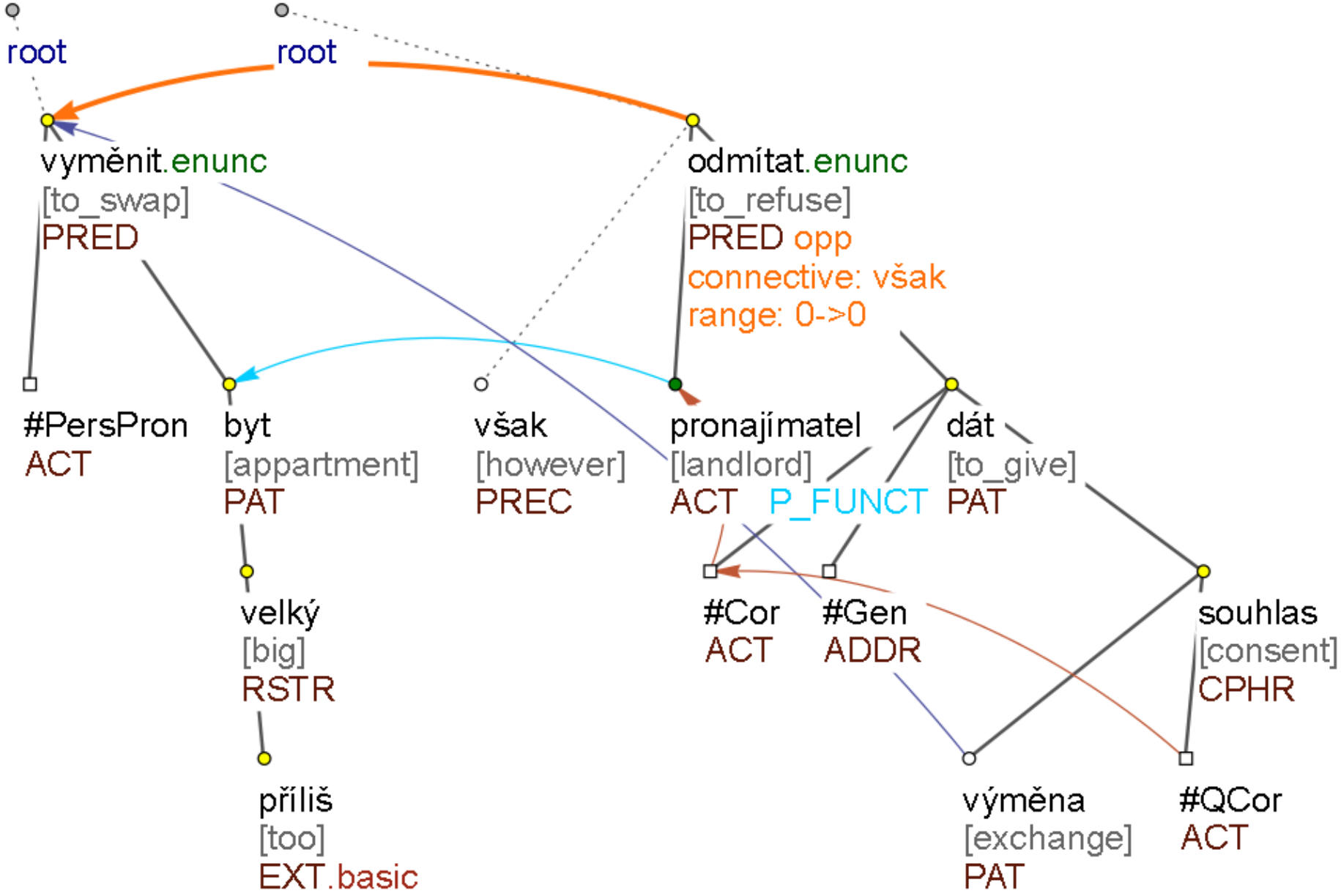

root
root
vyměnit.enunc
[to_swap]
PRED
odmítat.enunc
[to_refuse]
PRED
byt
[appartment]
PAT
však
[however]
PREC
pronajímatel
[landlord]
ACT
dát
[to_give]
PAT
velký
[big]
RSTR
souhlas
[consent]
CPHR
příliš
[too]
EXT.basic
výměna
[exchange]
PAT


root
root
vyměnit.enunc
[to_swap]
PRED
odmítat.enunc
[to_refuse]
PRED
#PersPron
ACT
byt
[appartment]
PAT
však
[however]
PREC
pronajímatel
[landlord]
ACT
dát
[to_give]
PAT
velký
[big]
RSTR
#Cor
ACT
#Gen
ADDR
souhlas
[consent]
CPHR
příliš
[too]
EXT.basic
výměna
[exchange]
PAT
#QCor
ACT


root
root
vyměnit.enunc
[to_swap]
PRED
odmítat.enunc
[to_refuse]
PRED
#PersPron
ACT
byt
[appartment]
PAT
však
[however]
PREC
pronajímatel
[landlord]
ACT
dát
[to_give]
PAT
velký
[big]
RSTR
#Cor
ACT
#Gen
ADDR
souhlas
[consent]
CPHR
příliš
[too]
EXT.basic
výměna
[exchange]
PAT
#QCor
ACT

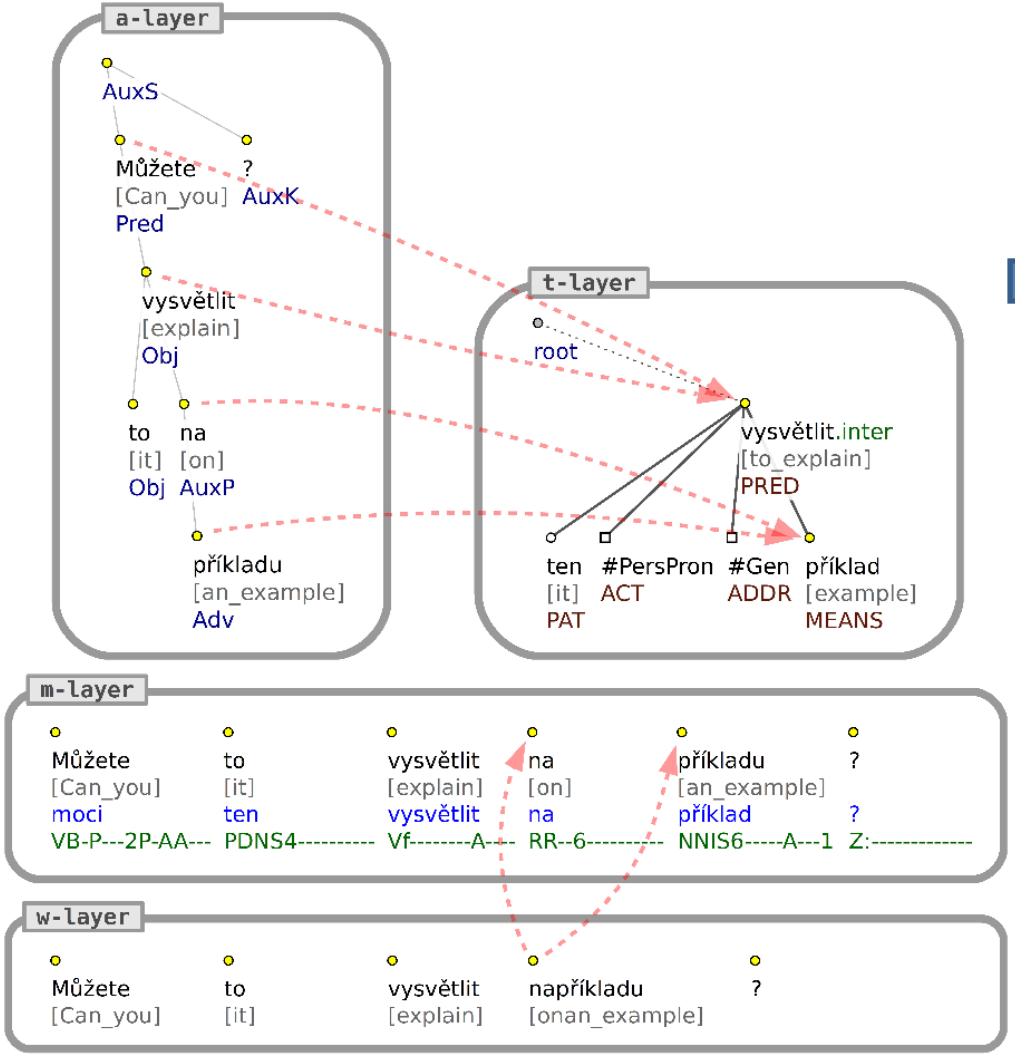

root
root
vyměnit.enunc
[to_swap]
PRED
odmítat.enunc
[to_refuse]
PRED
#PersPron
ACT
byt
[appartment]
PAT
však
[however]
PREC
pronajímatel
[landlord]
ACT
dát
[to_give]
PAT
velký
[big]
RSTR
#Cor
ACT
#Gen
ADDR
souhlas
[consent]
CPHR
příliš
[too]
EXT.basic
výměna
[exchange]
PAT
#QCor
ACT


root
root
vyměnit.enunc
[to_swap]
PRED
odmítat.enunc
[to_refuse]
PRED
#PersPron
ACT
byt
[appartment]
PAT
však
[however]
PREC
pronajímatel
[landlord]
ACT
P_FUNCT
dát
[to_give]
PAT
velký
[big]
RSTR
#Cor
ACT
#Gen
ADDR
souhlas
[consent]
CPHR
příliš
[too]
EXT.basic
výměna
[exchange]
PAT
#QCor
ACT


root
root
vyměnit.enunc
[to_swap]
PRED
odmítat.enunc
[to_refuse]
PRED opp
connective: však
range: 0->0
#PersPron
ACT
byt
[appartment]
PAT
však
[however]
PREC
pronajímatel
[landlord]
ACT
P_FUNCT
dát
[to_give]
PAT
velký
[big]
RSTR
#Cor
ACT
#Gen
ADDR
souhlas
[consent]
CPHR
příliš
[too]
EXT.basic
výměna
[exchange]
PAT
#QCor
ACT

Теоретическая база

Functional Generative Description (Sgall и др., 1986)



уровень дискурса?

хорошо, **НО**

смысл vs. значение

(анализ текста выходит за рамки языка)

План

I. Корпуса и дискурсивные явления

II. Пражский корпус – разметка

III. Исследования – прикладные, статистико-дескриптивные и лингвистические. Примеры.

   a. Пример 1: имплицитные анафорические отношения

   b. Пример 2: имплицитные дискурсивные отношения

IV. Выводы

Прикладные исследования

- Discourse parser (Mírovský et al.); Coreference parser (Novák et al.)
- Использование кореферентности и актуального членения для автоматического определения связности текста, в т.ч. на PDT (Novák, Mírovský, Rysová 2016, 2017)
- Автоматическая проекция кореферентности с одного языка на другой и анализ ошибок и расхождений (Novák, Nedoluzhko, Ogrodniczuk, Lapshinova 2016-2017)
- Создание словаря чешских дискурсивных коннекторов (Jinová, Mírovský, Rysová, Poláková)
- ....

Статистико-дескриптивные Исследования

- Квантитативные исследования и сравнение дискурсивных и кореферентных связей с подобными проектами (PDTB, GECCo, RuCor и др.)
- Анализ расхождений между разметчиками (в основном для дискурса и кореферентности) → вопрос множественности интерпретаций
- Взаимосвязь кореферентности и эллипсиса в PDT (Rysová&Rysová 2016)
- Исследования длины и устройства кореферентных цепочек, в т.ч. в сравнении с другими проектами (Недолужко-Толдова 2016)

Лингвистические исследования

- наличие ОТДЕЛЬНЫХ разметок текстовых явлений позвоняет глубже проникнуть в структуру текста

Лингвистические исследования

- наличие ОТДЕЛЬНЫХ разметок текстовых явлений позвоняет глубже проникнуть в структуру текста

Лингвистические исследования

- наличие ОТДЕЛЬНЫХ разметок текстовых явлений позвоняет глубже проникнуть в структуру текста
- мы можем рассматривать каждое явление в отдельности (с чего мы и начинали и что долго делали)
- рассмотрев и разметив каждое явление в отдельности, мы можем начать их постепенно сопоставлять и искать «**паттерны связности**»

Лингвистические исследования

- связь актуального членения и кореферентности:
  - кто такие некореферентные ИГ в теме?
  - какие значения АЧ имеют имеют узлы с анафорической отсылкой?
  - salience – расчет активированности с учетом кореферентных цепочек и активированности референта в тексте
- разрывы связного текста?
  - Что значит, если два предложения ничем «нашим» не связаны?
  - имплицитные анафорические отношения как средство связности текста
  - имплицитные дискурсивные отношения – чем отличаются от не-отношений? как их найт и
- сравнительное исследование поведения возвратно-поссессевного местоимения «свой»

Имплицитные Анафорические Отношения как Средство Когезии

- Выделение паттернов
- Можно выделить некоторые структуры с небазовым порядком слов, которые служат средством связности текста
- Рассмотрим конструкцию со связкой «быть»

Daniel je učitel.

Дан-NOM есть учитель-NOM.

Дан - учитель.

Daniel je učitelem.

Дан-NOM есть учитель -INSTR.

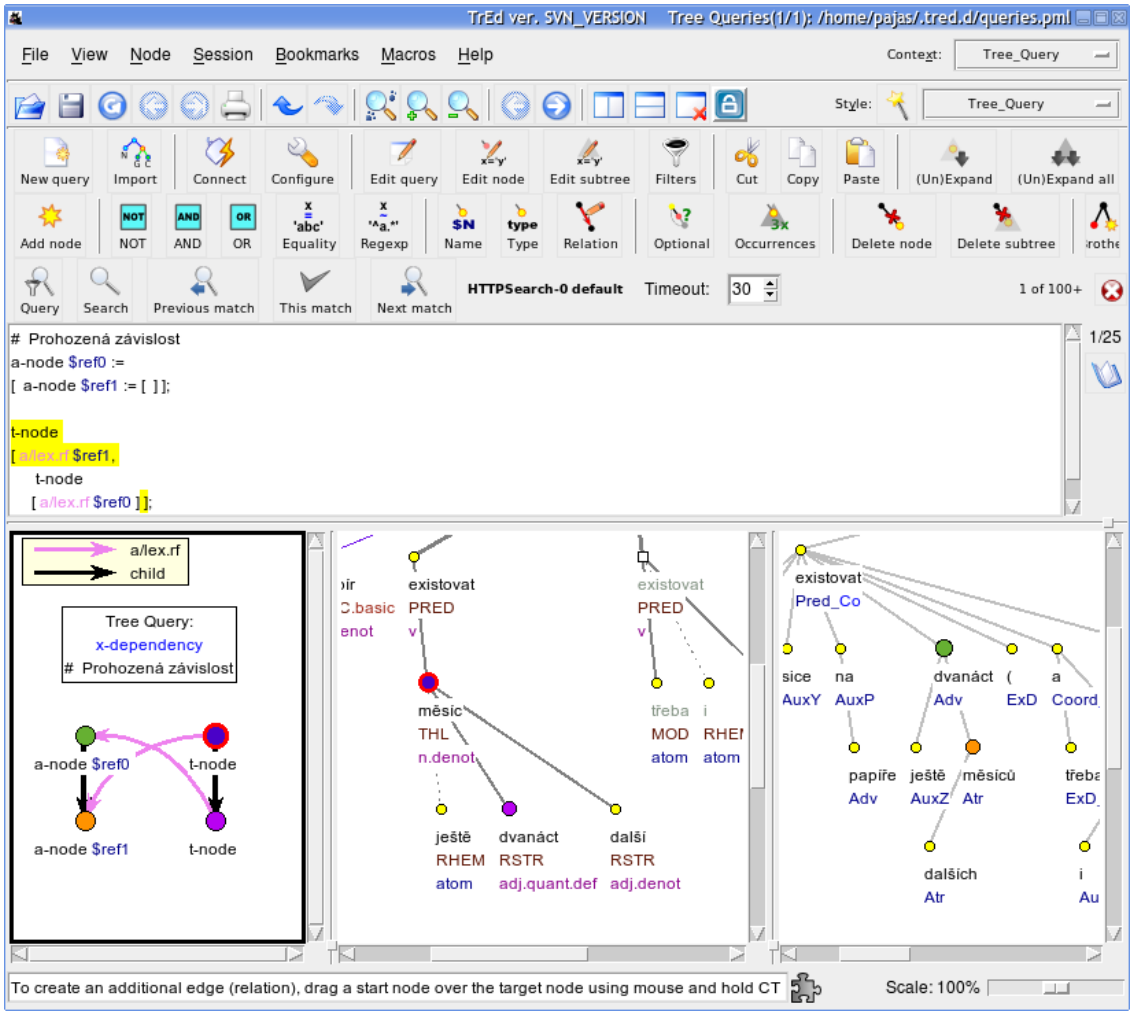
V Praze proběhne setkání mladých vědců.

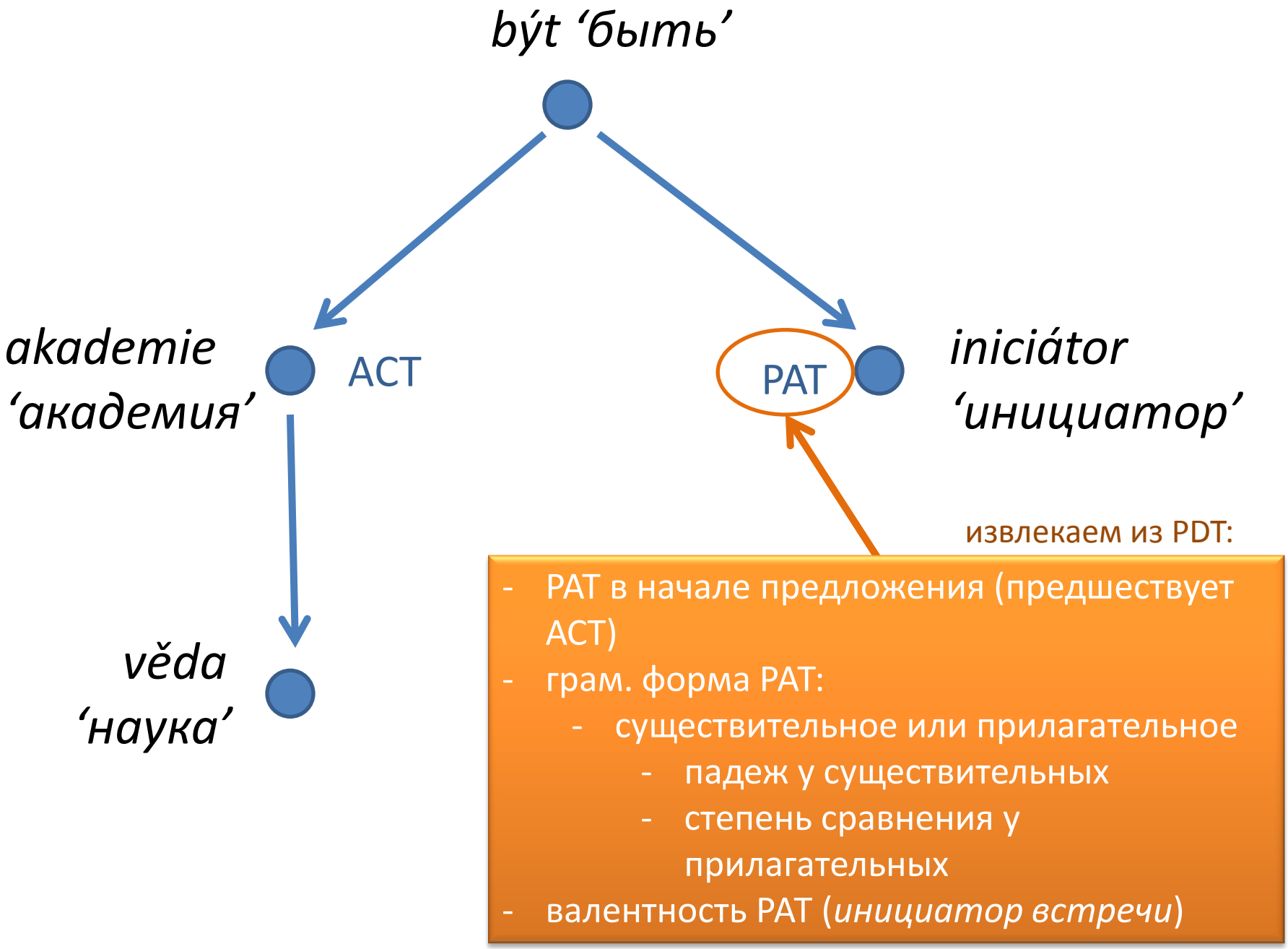
Iniciátorem -INST *je Akademie věd.*

В Праге пройдет встреча молодых ученых.
Инициатором является академия наук.

PML Tree Query

Search tool for { complex dependency constituency multi-layered parallel } treebanks




být ‘быть’
akademie
‘академия’
ACT
PAT
iniciátor
‘инициатор’
věda
‘наука’
извлекаем из PDT:
- PAT в начале предложения (предшествует ACT)
- грам. форма PAT:
  - существительное или прилагательное
    - падеж у существительных
    - степень сравнения у прилагательных
- валентность PAT (*инициатор встречи*)

Статистика из PDT

	Instrumental	**Nominative**	**Всего**
PAT initial (небазовый порядок слов)	529	243	772
ACT initial (базовый порядок слов)	474	1,718	2,192
Всего			2,964

Наблюдения

семантика

Две семантические группы PAT:

a. PAT обозначает различные аспекты антецедента (причинные отношения, пример, уступка и др.)

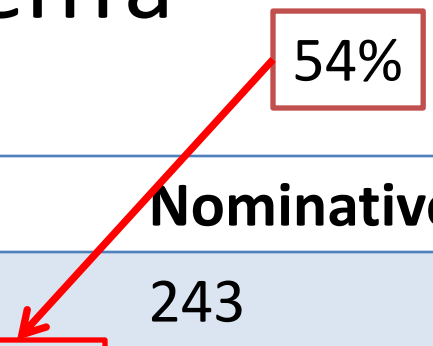
výsledkem / cílem/ příčinou / příkladem, … je

‘результатом/целью/причиной/примером,… было … ’

Tato fakta hovoří o nehotovosti českého politického systému.
Příkladem_INSTR *je situace v nejsilnější politické straně.*

Эти факты говорят о неподготовленности чешской политической системы. ***Примером*** *может служить ситуация в самой сильной политической партии.*

Статистика на PDT– аспекты антецедента



	Instrumental		Nominative		всего
PAT initial (небазовый порядок слов)	529		243		772
	Noun	526	Noun	14	
	Adj. positive	1	Adj. positive	172	
	Adj. comparative	0	Adj. comparative	37	
	Adj. superlative	2	Adj. superlative	20	
ACT initial (базовый порядок слов)	474		1,718		2,192
	Noun	469	Noun	552	
	Adj. positive	3	Adj. positive	979	
	Adj. comparative	0	Adj. comparative	171	
	Adj. superlative	2	Adj. superlative	16	
Всего					2,964

Наблюдения

семантика

Две семантические группы PAT:

а. PAT обозначает различные аспекты антецедента (причинные отношения, пример, уступка и др.)

б. PAT имеет семантику оценки

Dvořákův Jakobín neztratil svou hodnotu.
***Zklamáním*_INSTR.noun** *jsou však s odstupem let pěvecké výkony.*

Дворжаковский Якобинец не потерял своей ценности.
***Разочарованием**, однако, являются оперные партии.*

Наблюдения

семантика

Две семантические группы PAT:

а. PAT обозначает различные аспекты антецедента (причинные отношения, пример, уступка и др.)

б. PAT имеет семантику оценки

Trenér Gudlaugsson má v nominaci hráče podle svých představ. ***Nejzkušenějším*_INSTR.adj.superlative** *je obránce Jakobsen (31 let, 46x v reprezentaci).*

Игроки в номинации тренера Гудлаугсона соответствуют его ожиданиям. ***Самым опытным*** *является защитник Якобсен (31 год, 46 раз в репр.)*

Статистика на PDT – семантика оценки у PAT

45%

	Instrumental		Nominative		всего
PAT initial (небазовый порядок слов)	529		243		772
	Noun	526	Noun	14	
	Adj. positive	1	Adj. positive	172	
	Adj. comparative	0	Adj. comparative	37	
	Adj. superlative	2	Adj. superlative	20	
ACT initial (базовый порядок слов)	474		1,718		2,192
	Noun	469	Noun	552	
	Adj. positive	3	Adj. positive	979	
	Adj. comparative	0	Adj. comparative	171	
	Adj. superlative	2	Adj. superlative	16	
Всего					2,964

Статистика из PDT – сравнение падежей

	Instrumental		**Nominative**		**Всего**
PAT initial (небазовый порядок слов)	529		243		772
	Noun	526	Noun	14	
	Adj. positive	1	Adj. positive	172	
	Adj. comparative	0	Adj. comparative	37	
	Adj. superlative	2	Adj. superlative	20	
ACT initial (базовый порядок слов)	474		1,718		2,192
	Noun	469	Noun	552	
	Adj. positive	3	Adj. positive	979	
	Adj. comparative	0	Adj. comparative	171	
	Adj. superlative	2	Adj. superlative	16	
Всего					2,964

ИГ могут иметь внутреннюю структуру // валентность

(*начало* – это всегда начало чего-нибудь)

- лексическая (*отец* ↔ *тапочки*)
- грамматическая: у сравнительной степени прилагательного и у суперлативов (*лучше чего-то, самый лучший из кого-то*)

Наблюдения

синтаксис

- PAT (ИГ, adj.) с валентностью, где аргумент анафорически отсылает к предшествующему контексту имеют тенденцию употребляться в начале предложения
- PAT (ИГ, adj.) без валентности имеют тенденцию употребляться НЕ в начале предлоежениия

Наблюдения

синтаксис

Nehoda měla tragické následky. ***Příčinou*** *byla rychlá jízda.*

‘Авария имела трагические последствия. ***Причиной была быстрая езда****.’*

аварии

Эти анафорические отсылки...

- часто опустаются в начале предложения
- плохо опускаются в других позициях

Exploze na linii ropovodu byla ***příčinou*** *zastavení dodávek ruské ropy na Ukrajinu.*

Взрыв на нефтепроводе стал причиной остановки поставки российской нефти в Россию и на Украину.

анаф.ссылка опущена: 40%
анаф.ссылка выражена: 60 %

из них более половины может быть опущено

	Instrumental		Nominative		Всего
PAT initial (небазовый порядок слов)	529		243		772
	Noun	526	Noun	14	
	Adj. positive	1	Adj. positive	172	
	Adj. comparative	0	Adj. comparative	37	
	Adj. superlative	2	Adj. superlative	20	
ACT initial (базовый порядок слов)	474		1,718		2,192
	Noun	469	Noun	552	
	Adj. positive	3	Adj. positive	979	
	Adj. comparative	0	Adj. comparative	171	
	Adj. superlative	2	Adj. superlative	16	
Всего					2,964

Выводы:

- Можно выделить 2 семантические группы PAT в начале предложения: аспекты антецедента и оценка
- Есть взаимосвязь между синтаксической структурой и значением PAT в начале предложения: Instr. – больше каузальности, Nom. – больше оценки
- Валентность и невыраженные анафорические отношения: PAT в начале предложения часто имеет валентность анафорически связанную с предшествующим контекстом.

эти структуры систематически выражают связность текста

Лингвистические исследования

- связь актуального членения и кореферентности:
  - кто такие некореферентные ИГ в теме?
  - какие значения АЧ имеют имеют узлы с анафорической отсылкой?
  - salience – расчет активированности с учетом кореферентных цепочек и активированности референта в тексте
- разрывы связного текста?
  - Что значит, если два предложения ничем «нашим» не связаны?
  - имплицитные анафорические отношения как средство связности текста
  - имплицитные дискурсивные отношения – чем отличаются от не-отношений? как их найти?
- сравнительное исследование поведения возвратно-поссессевного местоимения «свой»

Имплицитные отношения

- Ищем паттерны, сигнализирующие наличие имплицитных отношений различного типа (Zikánová 2017, 2017a)
- Смотрим на то, чем имплицитные отношения отличаются от отсутствия отношения.

Имплицитные отношения

- Когерентность основана на кореференции:
  – идентичность субъектов в начале ДЕ
- Негация и коррекция
- Вводные конструкции и спецификация
- Оценка в первом аргументе и экспликация во втором аргументе

Негация и коррекция

Chtěl bych jasně říci, že to není podpora intenzivní zemědělské výroby.

CORRECTION

Je to vyjádření veřejného zájmu státu.

Я хотел быть подчеркнуть, что это не является поддержкой интенсивной сельскохозяйственной продукции.

CORRECTION

Это выражение публичного интереса государства.

Вводные конструкции и спецификация

- Подлежащее в финальной позиции (в фокусе)
- Следует спецификация

Celý tento orchestr řídí trenér Antonín Juran, expert na postupy.

SPECIFICATION

Jako hráč pomohl Zlínu vybojovat první ligu, pak přestoupil do Slovanu Bratislava.

Всем этим оркестром руководит тренер Антонин Юран, эксперт продвижения.

SPECIFICATION

Игроком он помог Злину выйти в первую лигу, потом он перешел в команду Слован в Братиславе.

Оценки и пояснение

- Первый аргумен содержит оценочную конструкцию
- Следует explication (justification)

Trh s těmito právy existuje v té nejhorší pololegální a nelegální podobě.

EXPLICATION

Nájemníci byty vyměňují, nelegálně pronajímají.

Торговля этими правами существует в своем самом ужасном пололегальном и нелегальном виде.

EXPLICATION

Владельцы меняют квартиры, нелегально их сдают.

Нам не хватает:

- Полярность (будет)
- Лексические вещи (оценочные выражения)
- Референциальный статус ИГ

План

I. Корпуса и дискурсивные явления
II. Пражский корпус – разметка
III. Исследования – прикладные, статистико-дескриптивные и лингвистические. Примеры.
    a. Пример 1: имплицитные анафорические отношения
    b. Пример 2: имплицитные дискурсивные отношения
IV. Выводы



Выводы:

- Корпус как тестовые и обучающие данные для автоматических программ
- Корпус – обучающие данные не только для программ
- Корпус дает идеи и материал для их первичной проверки
- Паттерны текстовой связности, которые мы найдем с помощью нашей интуиции основанной на корпусных данных, могут использоваться дальше для машинного обучения. И проверяться на данных другого масштаба