

Уровень текста в корпусной лингвистике: Пражский подход

Textual phenomena in the Prague Dependency Treebank

Анна Недолужко,
Москва, 22 мая 2015

* План

* Общие слова

* Люди и теории

* Дискурс, Textlink

* Кореферентность

* Пражский корпус в их контексте

* Текстовая разметка (кореферентность, бриджинг, эллипсис, актуальное членение, дискурс)

* Взаимодействие этих явлений между собой

* Два подхода: пражский и саарбрюкенский

Timestamp	Corpus name	Corpus acronym	Contact person e-mail(s)	Project URL	Availability	Languages covered	Available translations	Corpus ID
11.20.2014 9:22:11	Diff-En	Diff-En	ludvine.crible@univ-lorraine.fr	http://diff-en.univ-lorraine.fr/	still in progress of being a English, French	English, French	the corpus is comparable	N/A
11.21.2014 14:51:07	HiCoTech Multimodal CHM	HiCoTech	abuczki.agnes@gmail.com	https://hi-co-tech.univ-lorraine.fr/	Available for Academic - Hungarian	French		
11.27.2014 16:39:05	French Discourse Treebank FDTB1	FDTB1	Laurence.Daquin@univ-lorraine.fr	https://fdtb-v1.gforge.inria.fr/		English-to-German translation	Penn TreeBank corpus (v1 - 24)	
12.5.2014 11:29:27	Corpus for the Analysis of GECCO	ANNODS	e.lapshinova@univ-lorraine.fr	http://www.univ-lorraine.fr/gecco/	it will be soon publicly (re)English, German			
12.7.2014 18:59:16	Penn Discourse Treebank PDTB 2.0	ANNODS	bonnie.welsh@ed.ac.uk	http://www.ed.ac.uk/pdtk/	Contact the Linguistics D-English			
12.9.2014 16:35:57	ANNODS	ANNODS	hustad@univ-lorraine.fr	http://www.univ-lorraine.fr/creative-commons/	By request	French		
12.10.2014 11:55:08	French PropBank	ANNODS	veronica.lapalme@univ-lorraine.fr	https://github.com/veronica-lapalme/french-propbank	Publicly available in http://Greek	French		
12.10.2014 15:55:40	CSTNews	CSTNews	lesbeth.dagan@univ-lorraine.fr	http://www.univ-lorraine.fr/cstnews/	Publicly available in http://Greek	French		
12.10.2014 16:17:50	Loirean Corpus of Aeneas LOCAS-F	LOCAS-F	antigoni.papamargarita@univ-lorraine.fr	http://www.univ-lorraine.fr/locas-f/	Downloaded from webpage	German		
12.11.2014 10:02:41	Hellenic National Corpus HNC	HNC	stedes@univ-lorraine.fr	http://www.univ-lorraine.fr/hnc/	publicly available in UNL/Czech	Hebrew		
12.11.2014 10:36:23	Corpus of Greek Texts CGT	CGT	maschke@univ-lorraine.fr	http://www.univ-lorraine.fr/cgt/	not available	Italian		
12.13.2014 14:49:32	Prague Dependency Treebank PDTB 3.0	PDTB 3.0	matthew@univ-lorraine.fr	http://www.univ-lorraine.fr/pdtb3/	Currently negotiating with Catalan	Turkish		
12.14.2014 14:47:55	The Halls Corpus of Spoken HALLS	HALLS	ludvine.crible@univ-lorraine.fr	http://www.univ-lorraine.fr/halls/	Upon request	EL		
12.19.2014 12:17:24	Prague Dependency Treebank PDTB 3.0	PDTB 3.0	reza.saei@gmail.com	http://www.univ-lorraine.fr/pdtb3/	Upon request via https://ic/Dutch			
12.22.2014 21:36:35	Turkish Discourse Bank TDB	TDB	voulas@univ-lorraine.fr	http://www.univ-lorraine.fr/tdb/	Upon request via https://ic/Dutch			
12.30.2014 16:03:17	LUNA corpus	LUNA	l.j.m.sander@univ-lorraine.fr	http://www.univ-lorraine.fr/luna/	Linguistic Data Consortium English	English, Spanish, Basque/Parts in French		
1.9.2015 8:27:03	Catalan Discourse Bank CDB	CDB	mtaboad@univ-lorraine.fr	http://www.univ-lorraine.fr/cdb/	publicly available at http://Basque			
4.15.2015 10:11:34	Greek Sentiment Corpus GSC-Oral	GSC-Oral	mikel.nuskiet@univ-lorraine.fr	http://www.univ-lorraine.fr/gsc-oral/				
4.15.2015 11:22:22	DiscAn - Towards a Discourse Analysis	DiscAn	mikel.nuskiet@univ-lorraine.fr	http://www.univ-lorraine.fr/discan/				
4.20.2015 23:01:25	RST Signaling Corpus RST-SC	RST-SC						
6.7.2015 1:40:26	The Basque RST Treebank Basque RST-TB	RST-TB						
6.7.2015 1:52:35	The Multilingual RST Tree Multilingual RST-TB	RST-TB						

precedence - succession	reason - result	confrontation	conjunction
synchronous	pragmatic reason - result	opposition	instantiation
	purpose	pragmatic contrast	specificity
	explication	restrictive opposition - exception	equivalence
	condition	concession	generalization
	pragmatic condition	correction (replacement)	conjunctive alternative
		gradation	disjunctive alternative

TEMPORAL	CAUSAL	ADVERSATIVE	ADDITIVE	MODAL
temporal relation between events	relation of causality/dependence between	relation of contrast/alternation, for two events which are not true at the same time	relation of addition, for two events which are true at the same time	relation between two events connected by an arrow, of the speaker
after, afterwards, at the same time...	because, therefore, that's why...	yet, although, by contrast...	and, in addition...	well, sure, of course, eventually...
nachdem, danach, gleichzeitig...	weil, deshalb, aus diesem Grund...	doch, obwohl, im Gegensatz dazu...	und, außerdem...	klar, sicher, allerdings, jedenfalls, irgendwann...

Текст

- * когерентность
- * когезия
- * коммуникативная цель
- * информативность
- * ситуативность (место, время)
- * интертекстовость
- * др.

Dressler, W. (1972)

- актуальное членение, тематические цепочки;
- кореферентность, кореф.цепочки; анафорические отношения;
- временные взаимосвязи;
- графическая и звуковая организация текста;
- риторические отношения (семантические, рематические, композиционные);
- синтаксическая структура текста (+текстовые коннекторы, эллипсисы, диск. маркеры)

- коммуникативные и прагматические факторы;
- жанры,
- evaluating modality (sentiment, negation),
- разметка речевых актов (угроза, убеждение и др.)
- критический дискурс-анализ (CDA, Chouliaraki and Fairclough 1999) - как формируются стереотипы и их лингв. аспекты, дискурс в социальном плане

Текст

- * когерентность
- * когезия
- * коммуникативная цель
- * информативность
- * ситуативность (место, время)
- * интертекстовость
- * др.

Dressler, W. (1972)

- актуальное членение, тематические цепочки;
- кореферентность, кореф. цепочки; анафорические отношения;
- временные взаимосвязи;
- графическая и звуковая организация текста;
- риторические отношения (семантические, рематические, композиционные);
- синтаксическая структура текста (+текстовые коннекторы, , эллипсисы, диск. маркеры);

- коммуникативные и прагматические факторы;
- жанры,
- evaluating modality (sentiment, negation),
- разметка речевых актов (угроза, убеждение и др.)
- критический дискурс-анализ (CDA, Chouliaraki and Fairclough 1999) - как формируются стереотипы и их лингв. аспекты, дискурс в социальном плане



Мотивация

- * Теоретическая

- * анализ текста

- * типология

- * неоднозначность

- * → понимание и интерпретация текста

- * NLP

- * машинный перевод

- * создание моделей общения, диалоговые системы

- * компьютерный анализ документов (классификация, поиск, суммаризация и т.д.)

- * автоматическая разметка новых данных ...

* Подходы и теории



* 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

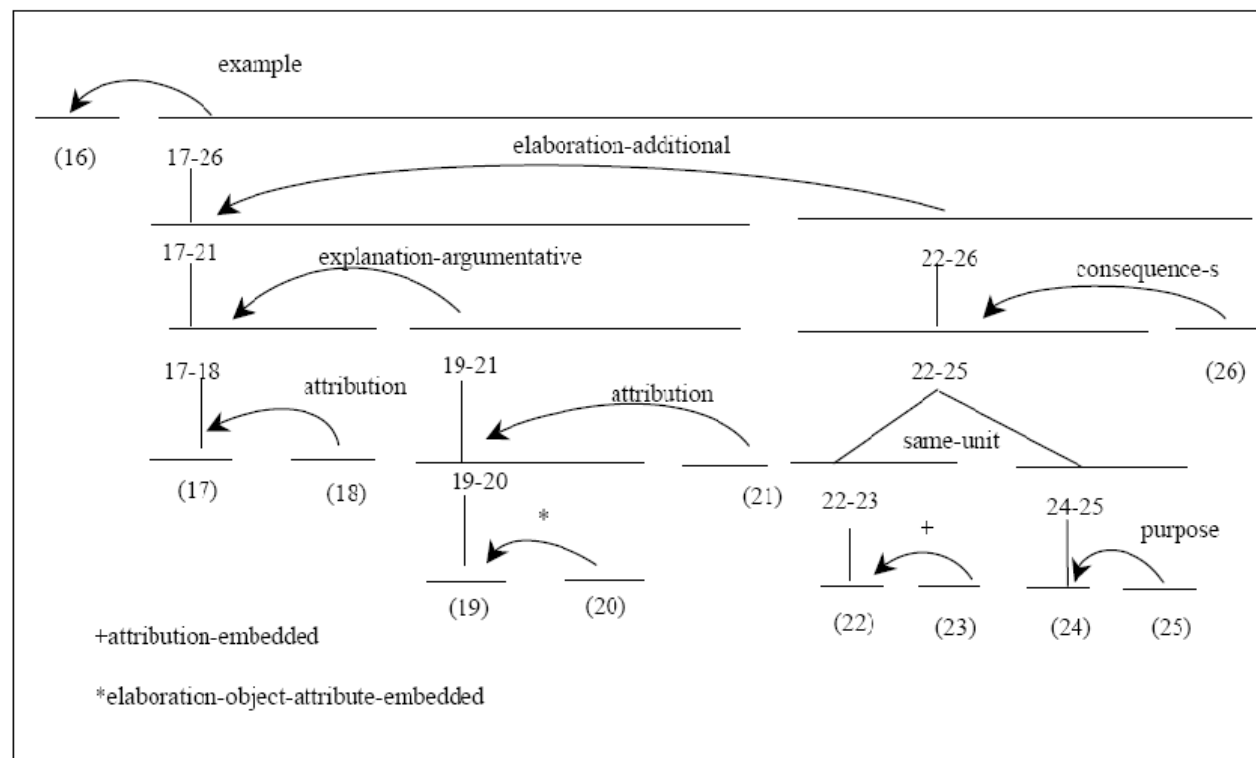
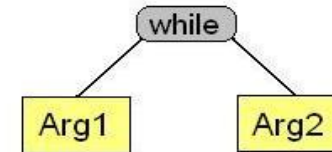


Figure 1. Discourse sub-tree for multiple sentences

*Penn Discourse TreeBank 2.0

John eats porridge for breakfast, while 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 spokeswoman 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

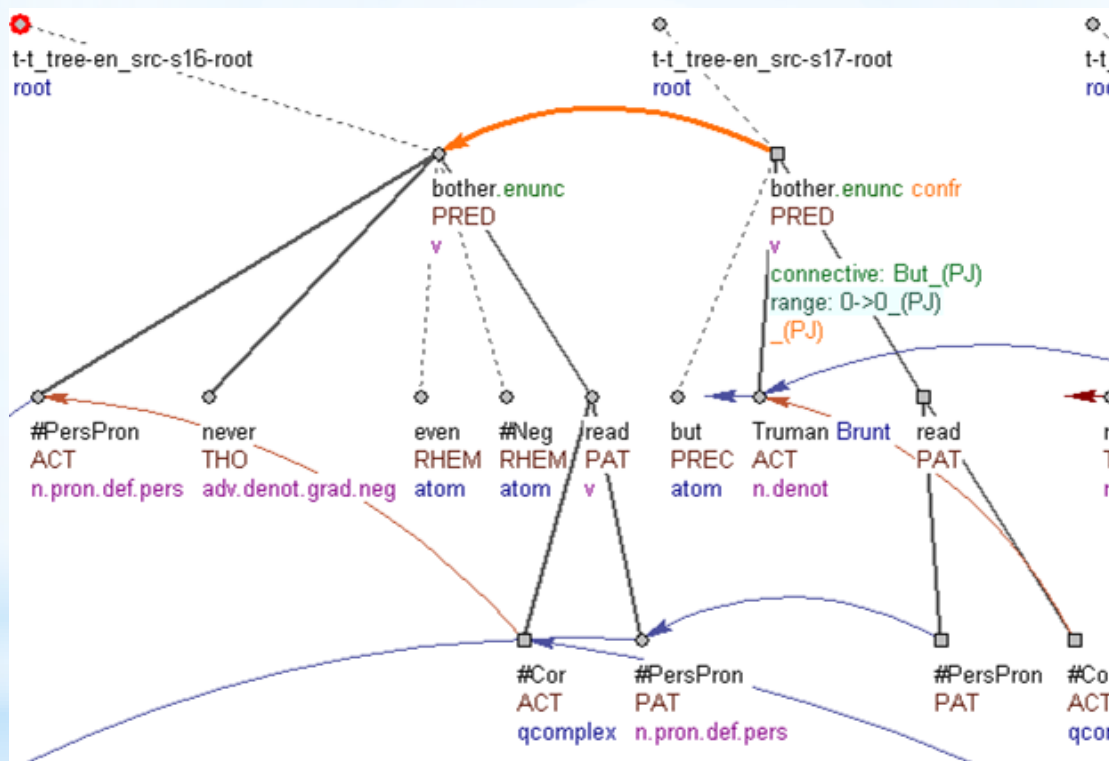
* German-English Contrasts in Cohesion (Kunz et al. 2014)

far out on the frozen tip of the continent even the polar bears couldn't find him ? Betr
ie title again with a slow studied movement of his lips . It was only words , only history .
s . He ' d never even bothered to read it . But Truman had . At the moment Truman wa
air of unconcern , as if showing his alienated son the work of his mad wasted life was ar
ll gin and lemonade that he was wrought up . The old man suddenly darted a glance ov
hering revelation , his father the phantom come to life , the book of the dead spread o

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

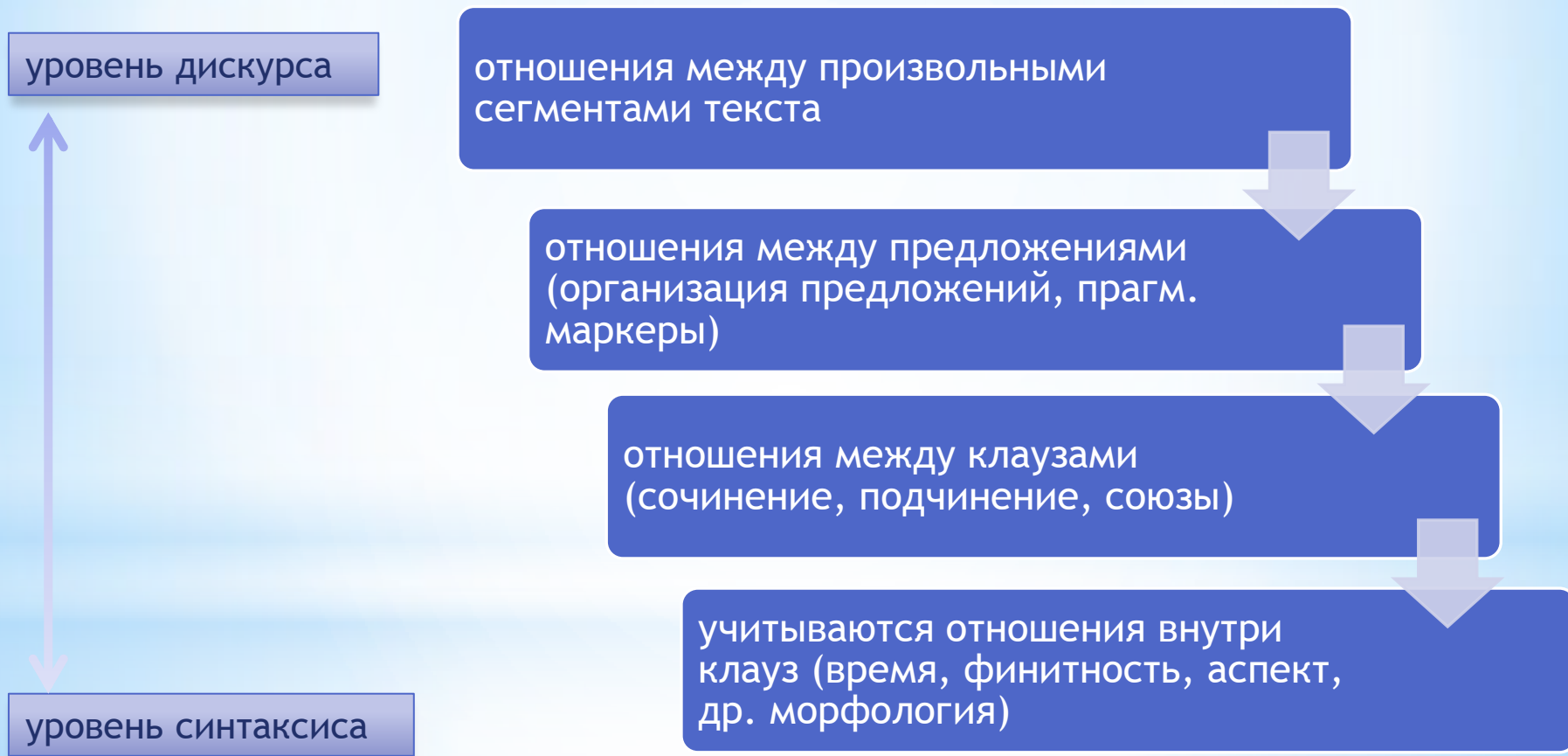
reference	conj	ellipsis	substitution
type	connect ▼		
func	adversative ▼		
problematic	☒ no ☐ yes		
☒ Suppress check ☒ Warn on extra attributes			

* Prague Dependency Treebank (Poláková et al. 2014)



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

* Дискурсивная разметка: подходы и теории



* Дискурсивная разметка: подходы и теории - *draft*



* Граница знаний о мире и внутритекстовых связей

- * Психолингвистические эксперименты Келлера (Kehler - Rohde, 2013)
- * Cohesion-Driven x Centering-Driven Theories: интерпретации местоимений - сравнимы
- * Трудноотделимость нетекстового от текстового

* Дискурсивные разметки для других языков:



- * китайский (Zhou - Xue, 2012),
- * арабский (Al-Saifová - Markertová, 2010),
- * турецкий (Zeyreková и др., 2010),
- * хинди (Oza и др., 2009),
- * французский (Afantenos и др., 2012, Danlos и др., 2012),
- * немецкий (Stede - Neumann, 2014; Gastelová и др., 2011),
- * чешский (Poláková и др. 2013)
- * венгерский (Abuczki)
- * финский (Finnish PropBank, Laippala)
- * португальский (CSTNews, Mendez)
- * греческий, иврит, итальянский, каталанский, датский, баскский и др.



ISCH OOST Action IS1312




[Home](#)
[Documents](#)
[Management Committee](#)
[Working Groups](#)
[Networking Tools](#)
[Communication](#)

Working Groups

WG1: Resources

This WG is in charge of assembling a list of existing corpora and DRD inventories in various languages, checking copyright issues, and developing a systematic description of each in the form of standardised metadata (to support Interoperable search and facilitate comparisons between languages, genres, modes, etc.).

[More](#)

WG2: Interoperable Annotation Guidelines

Based on a detailed study and comparison of both theoretical accounts and practical applications in already annotated corpora, this WG will work towards an interoperable conceptual framework for the multilingual annotation of the meanings conveyed by DRDs across European and non-European languages.

[More](#)

WG3: Assessment of Empirical and Cognitive Soundness

This WG approaches DRDs from the perspectives of methodologically-sound cross-linguistic analysis, and of psycholinguistic experimentation. Within the framework of the Action "best practices" will be confronted, including results on inter-annotator agreement, compatibility of the taxonomies with psycholinguistic findings, and performance on automatic sense annotation.

[More](#)

WG4: Tools

The WG will make recommendations for multi-layer annotation tools that support effective annotation of DRDs and relations across languages, tools for automating parts of the process to enable more efficient use of human annotators, and tools that can use the assigned metadata to integrate, and enable search over multiple DRD annotated corpora.

[More](#)

Проект TextLink:

Structuring Discourse in Multilingual Europe

	A	B	C	G	H	I	J	K	
	Timestamp	Corpus name	Corpus acronym	Contact person e-mail(s)	Project URL	Availability	Languages covered	Available translations	Corp
1	11.20.2014 9:22:11	DisFrEn	DisFrEn	ludvine.crible@uclouvain.be	http://uclouvain.be/435162	still in progress of being a	English, French	the corpus is comparable	
3	11.21.2014 14:51:07	HuComTech Multimodal CHMC		abuczki.agnes@gmail.com	https://hucomtech.unideb.hu/	Available for Academic - N	Hungarian	N/A	
4	11.27.2014 16:39:05	French Discourse Treebank	FDTB1	Laurence.Danlos@inria.fr	https://fdtb-v1.gforge.inria.fr/		French		
5	12.5.2014 11:29:27	Corpus for the Analysis of	GECCo	e.lapshinova@mx.uni-saarland.de	www.gecco.uni-saarland.de	it will be soon publicly (re-	English, German	English-to-German transla	604
6	12.7.2014 18:59:16	Penn Discourse TreeBank	PDTB 2.0	bonnie.webber@ed.ac.uk	http://www.seas.upenn.edu/	Contact the Linguistics D	English	Penn TreeBank corpus (or	~24
7	12.9.2014 16:35:57	ANNODIS	ANNODIS	hodac@univ-tlse2.fr	http://redac.univ-tlse2.fr/c/	Creative Commons By-NC	French		
8	12.10.2014 11:55:08	Finnish PropBank		veronika.laippala@utu.fi	bionlp.utu.fi		Finnish		
9	12.10.2014 14:45:58	Contexts of subordination		jyrki.kalliokoski@helsinki.fi	https://tuhat.helsinki.fi/portal/fi/projects/context		Finnish		
10	12.10.2014 15:55:40	CSTNews Corpus	CSTNews	amalia.mendes@clul.ul.pt	http://www.nilc.icmc.usp.br/	By request.	Brazilian Portuguese		
11	12.10.2014 18:17:50	Louvain Corpus of Annota	LOCAS-F	liesbeth.degand@uclouvain.be	N/A	N/A	French	N/A	N/A
12	12.11.2014 10:02:41	Hellenic National Corpus	HNC	antigoni.parmaxi@cut.ac.uk	http://hnc.ilsp.gr/en/	Publicly available in http://	Greek	Only the interface is trans	
13	12.11.2014 10:36:23	Corpus of Greek Texts	CGT	antigoni.parmaxi@cut.ac.uk	http://www.sek.edu.gr/	Publicly available at http://	Greek	Only the interface is available i	
14	12.13.2014 14:49:32	Potsdam Commentary Cc	PCC 2.0	stede@uni-potsdam.de	https://www.ling.uni-potsdam.de/	Download from webpage	German		
15	12.14.2014 14:47:55	The Haifa Corpus of Spok	N/A	maschler@research.haifa.ac.il	http://weblix2.haifa.ac.il/~by	request	Hebrew	N/A	325
16	12.19.2014 12:17:24	Prague Dependency Tree	PDT 3.0	mirovsky@ufal.mff.cuni.cz	https://ufal.mff.cuni.cz/pdf	publicly available in LINDA	Czech	none	
17	12.22.2014 16:00:17	LUNA corpus	LUNA	satonelli@fbk.eu		Not available	Italian		60 d
18	12.30.2014 21:36:35	Turkish Discourse Bank	TDB	ruken@ceng.metu.edu.tr	http://medid.ii.metu.edu.tr/	corpora@metu.edu.tr	Turkish	none	
19	1.9.2015 8:27:03	Catalan Discourse Graph	CatDiG	roser.sauri@gmail.com	N/A	Currently negotiating distr	Catalan	N/A	
20	4.15.2015 10:11:34	Greek Sentiment Corpus - GSeC-Oral		voula@ilsp.gr		Upon request	EL		
21	4.15.2015 11:27:22	DiscAn - Towards a Disc	DiscAn	t.j.m.sanders@uu.nl	https://tla.mpi.nl/resource	Upon request via https://c	Dutch	N/A	N/A
22									
23	4.20.2015 23:01:25	RST Signalling Corpus	RST-SC	mtaboada@sfu.ca	http://www.sfu.ca/~mtabo	Linguistic Data Consortiur	English		
24	5.7.2015 1:40:26	The Basque RST Treeban	Basque RST-TB	mikel.iruskieta@ehu.eus	http://ixa2.si.ehu.es/disk	publicly available at http://	Basque	The most of the text is av	
25	5.7.2015 1:52:35	The Multilingual RST Tree	Multilingual RST-TB	mikel.iruskieta@ehu.eus	http://ixa2.si.ehu.es/rst/	http://ixa2.si.ehu.es/rst/	English, Spanish, Basque	Parts in French	

*Кореферентность и анафора - теории и корпуса

- кореферентность в компьютерной лингвистике
 - ✓ детальная ручная разметка на больших корпусах: MATE (Poesio) – английский, PoCoS (английский, немецкий), VENEX (итальянский), ANCORA (испанский), OntoNotes (английский), Tutin (французский + др. языки).
 - ✓ большие корпуса, простые схемы, статистические методы, много языков: MUC (английский), ACE (англ., арабский, китайский ...)
- анафорические отношения в рамках теории референции
 - ✓ российские Е.В. Падучева, Н.Д. Арутюнова, А.Д. Шмелев ...
 - ✓ референциальный выбор (А.А.Кибрик)
 - ✓ чешские теории референции (Daneš, Palek, Hlavsa, Adamec и др.)
 - ✓ немецкие слависты (Berger, Mendoza)
 - ✓ др. мировые звезды в этой области (Carlson, Lyons, ...)
- когнитивные подходы (Consten, Knees, Schwarz-Friesel)

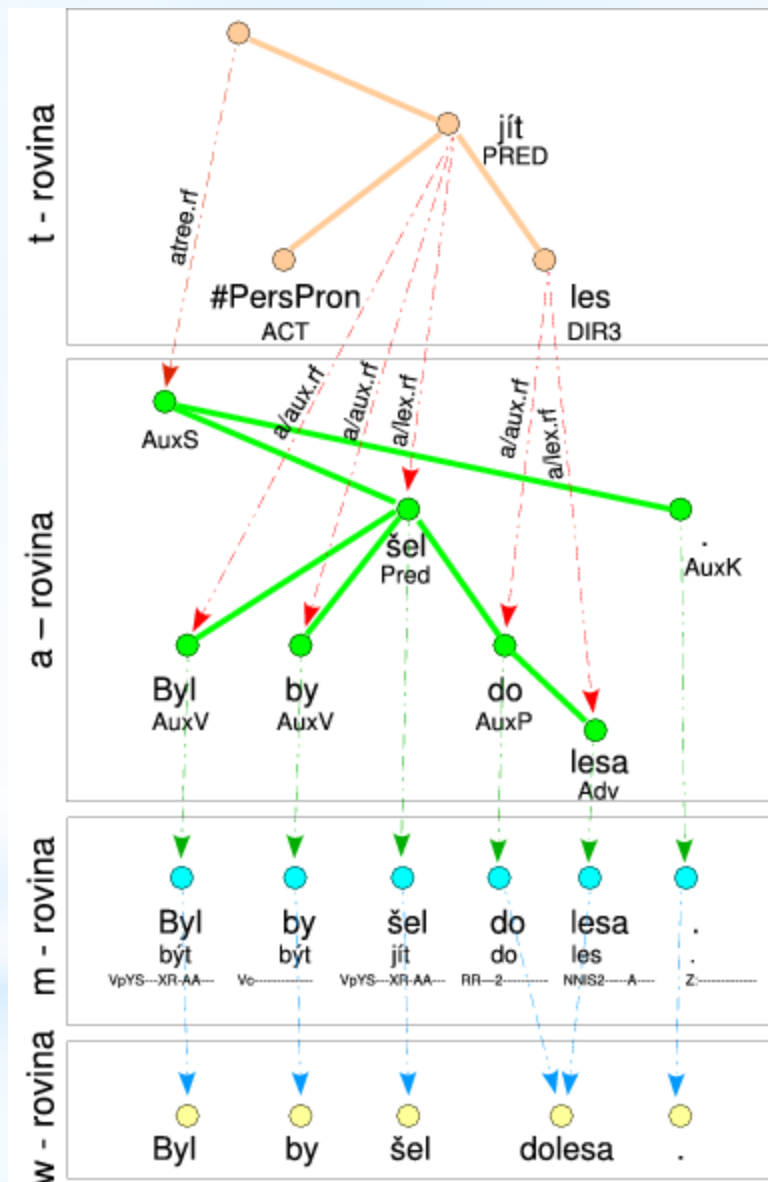
Prague Dependency Treebank (PDT 3.0)

Prague Dependency Treebank (PDT 3.0)

- * The Prague Dependency Treebank - синтаксически размеченный корпус, чешские газетные тексты (<http://ufal.mff.cuni.cz/pdt2.0>), данные доступны с LDC (No. LDC2006T01)
- * три уровня лингвистической разметки: морфологический, поверхностно-синтаксический (аналитический) и глубинно-синтаксический (тектограмматический)
- * 3165 текстов (= 49431 предложение = 833195 токенов)
- * многоцелевой
- * другие корпуса ÚFAL: <http://ufal.mff.cuni.cz/data>

*Byl by
šel
dolesa.*

*(Он)
пошёл
бы в лес.*



* Разметка на уровне текста

- * 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



* Наши корни



- * Functional Generative Description (Sgall и др., 1986)
- * кореферентность
 - * Ontonotes (Pradhan и др. 2007), Poesio (MATE, GNOME, VENEX, ARRAU), AnCora (Recasens), PoCoS (Chiarchos-Krasavina) и др., только мы это делаем на синтаксических деревьях
- * дискурсивная разметка
 - * Penn Discourse Treebank (для англ: PDTB, Prasad и др., 2008) - но мы чуть менее лексикализованы

*Текстовая разметка в пражских корпусах

	PDT 3.0	PCEDT (Prague Czech-English Dependency Treebank)	
		Английский	Чешский
синтаксическая и прономинальная (нули + мест.) кореферентность	ДА	ДА	ДА
Текстовая NP кореферентность	ДА	ДА	ДА
бриджинг	ДА	НЕТ(кроме split ante)	НЕТ(кроме split ante)
эллипсисы	ДА	ДА	ДА
коннекторы и отношения	ДА	НЕТ	НЕТ
актуальное членение	ДА	планируется	в процессе

*Кореферентность

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

* Грамматическая корелерентность



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

*Текстовая кореферентность

- * личные и possessивные 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 (рго) (... и гораздо больше рго 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=НЕ) (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 корелерентность - типы

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

(Ваня_а попросил маму_б объяснить_с ему_а, как ему_а вести себя с Машей, но мама_б сочла эту просьбу_с своего_б сына_а абсурдной)

- * GEN(eric) корелерентные отношения родовых и других нерелерентных 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.)

КОНЕЦ СКАЗКИ

Ассоциативная анафора (бриджинг)

Ассоциативная анафора (бриджинг)



- * некорреферентное 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 – участник (проституция - проститутка)

Статистика по корелерентности и бриджинг

Тип отношения	количество
Всего текстовой корелерентности	86,349
Текстовая корелер. (конкр-реф. NP)	20,243 (мест.)+50,593 (сущ.) = 70,836
Текстовая корелер. (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 (корелер. или бриджинг)	17.6%

Соответствие разметчиков (IAA)

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. карра на соответствие типов	0,759
bridging карра на соответствие типов	0,889

Обзор резолверов по корелерентности

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 резолверов по 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 резолверов по корелерентности

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

Нули

Нули с кореферентностью



t-lemma	описание	пример
#PersPron	Присваивается узлам, представляющим личные и possessивные местоимения - как выраженным, так и восстановленным	<i>(Firma měla doručit zboží zákazníkovi.) Doručení {#PersPron.ACT} {#PersPron.PAT} {#PersPron.ADDR} se však neuskutečnilo. (=lit. (Фирма должна была доставить товар заказчику.) Однако доставка {#PersPron.ACT} {#PersPron.PAT} {#PersPron.ADDR} не осуществилась.)</i>
#Cor	Присваивается восстановленным узлам - в контролирующих конструкциях	<i>Ольга решила {#Cor.ACT} ответить ему.</i>
#QCor	Присваивается восстановленным узлам - актантам в квазиконтролирующих конструкциях.	<i>Он предложил Ольге {#QCor} помощь.</i>
#Rcp	Присваивается восстановленным узлам - актантам в конструкциях с рецироком.	<i>Мальчики поцеловались {#Rcp.PAT}.</i>
Восстановленная лексема	В случае текстового эллипсиса, копия узла с такой же леммой восстанавливается на тектограмматическом уровне.	<i>Клиенты закрывшихся страховых компаний автоматически становятся клиентами центральной {страховой компании}.</i>

Нули без кореферентности



t-lemma	описание	пример
#Gen	Присваивается восстановленным узлам - общим (родовым) актантам	<i>Дома строятся {#Gen.ACT} из кирпичей.</i>
#Unsp	Присваивается восстановленным узлам - актантам с неясным (экзистенциальные, неконкретные, универсальные и т.д. по Падучевой)	<i>У Петровых {#Unsp.ACT} вкусно готовят.</i>
#EmpNoun	Присваивается восстановленным узлам - невыраженным существительным, управляющим выраженными прилагательными в случаях иных, чем текстовый эллипсис.	<i>{#EmpNoun.ACT} Проигравшие плакали.</i>

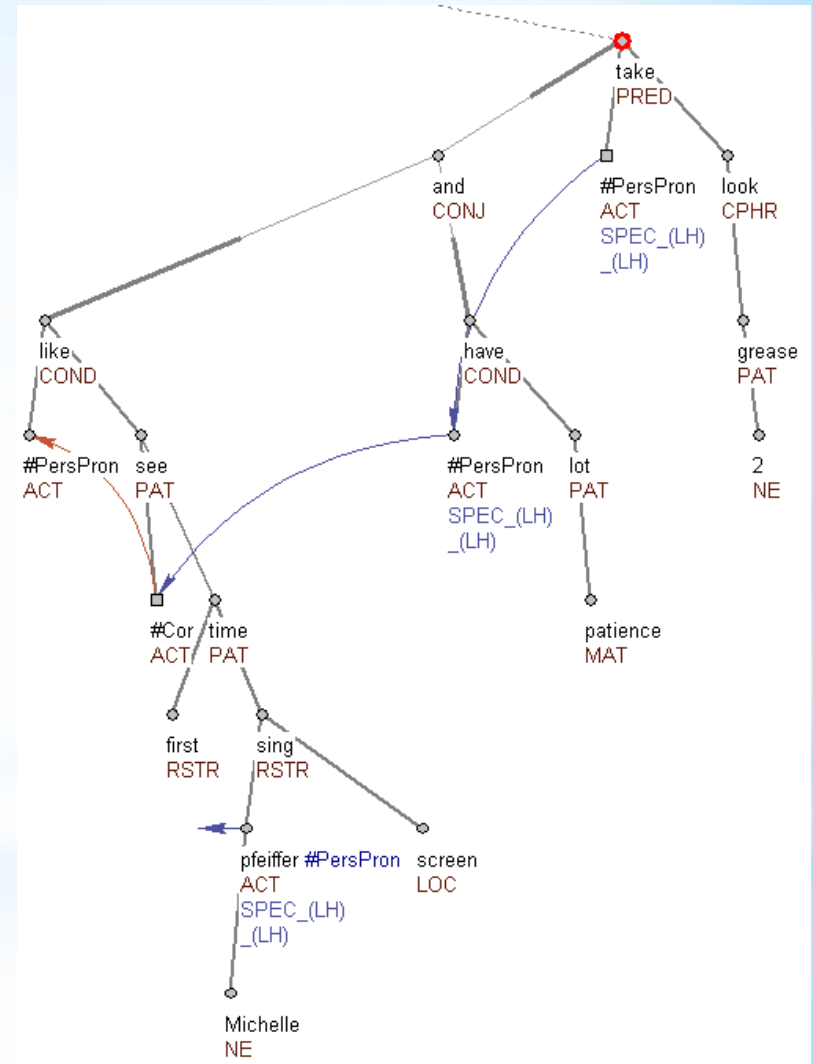
Нули без кореферентности



t-lemma	описание	пример
#Oblfm	Присваивается восстановленным узлам - обязательным адъюнктам.	Ta vypadá. {#Oblfm.MANN} (=lit. That.fem looks; meaning: She looks awful/so strange...)
#Equal	Присваивается восстановленным узлам в сравнительных конструкциях	Udělal to jako Tonda. (=lit. (He) did it like Tonda)
#Some	Присваивается восстановленным узлам - именная часть вербономинального предиката в сравнительных конструкциях	Není ovšem ford jako ford. - lit. However, Ford is not like Ford (meaning Not all Fords are the same)

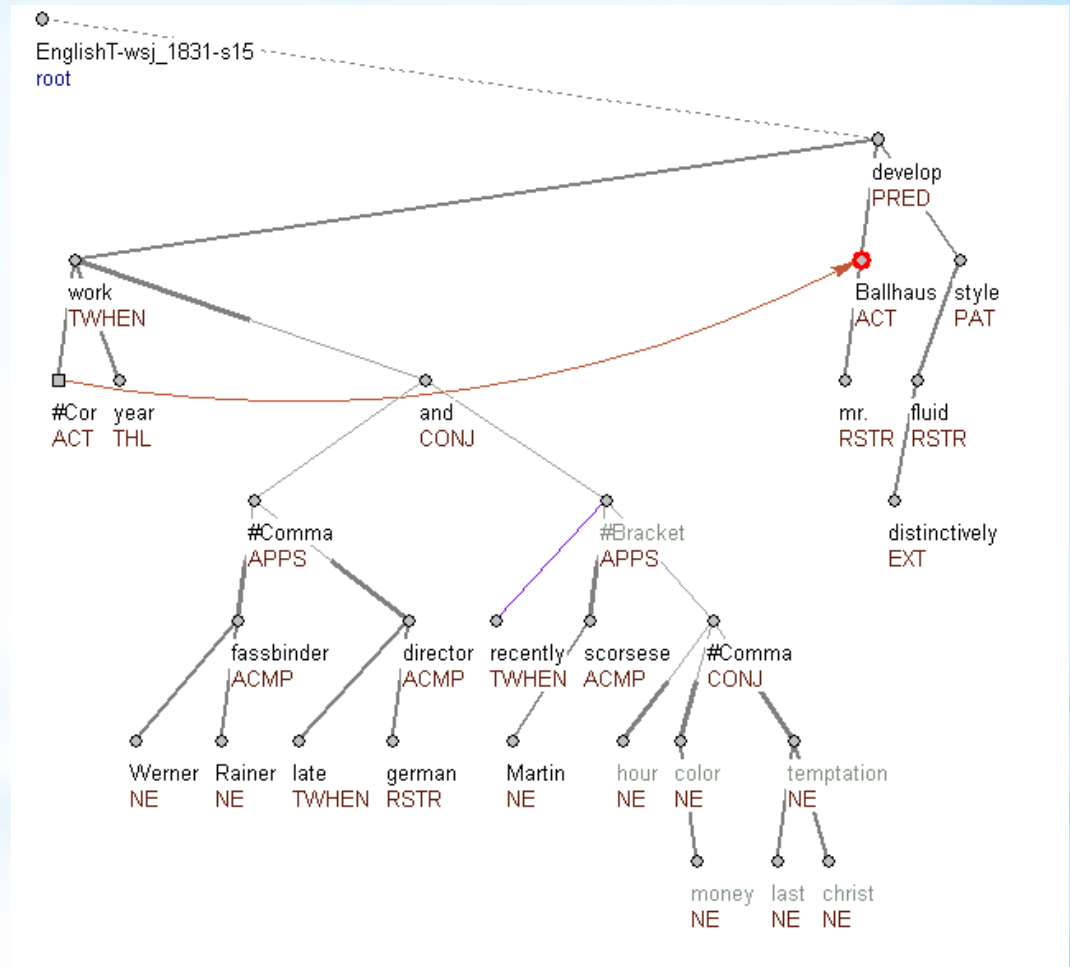
#PersPron, #Cor

If you'd like {#Cor.ACT} to see the first time Michelle Pfeiffer sang on screen, and you have a lot of patience, #PersPron take a look at “Grease 2”.



#Cor

After {#Cor.ACT} working for years with Werner Rainer Fassbinder, the late German director, and more recently with Martin Scorsese "After Hours", "The Color of Money", "The Last Temptation of Christ", Mr. Ballhaus has developed a distinctively fluid style.

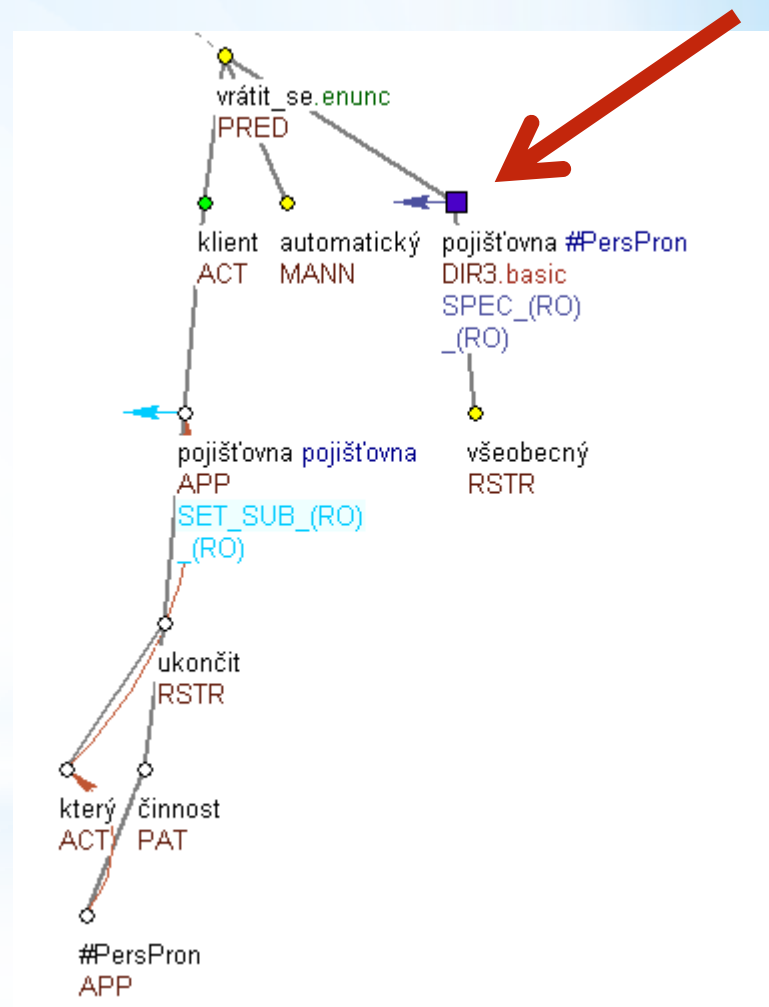


Восстановлено опущенное существительное

If it is clear (and possible to find in the text) which noun has been omitted in the surface structure of the sentence (the case of textual ellipsis), a copy of the node representing the same lexical unit as the omitted element is inserted into the appropriate position.

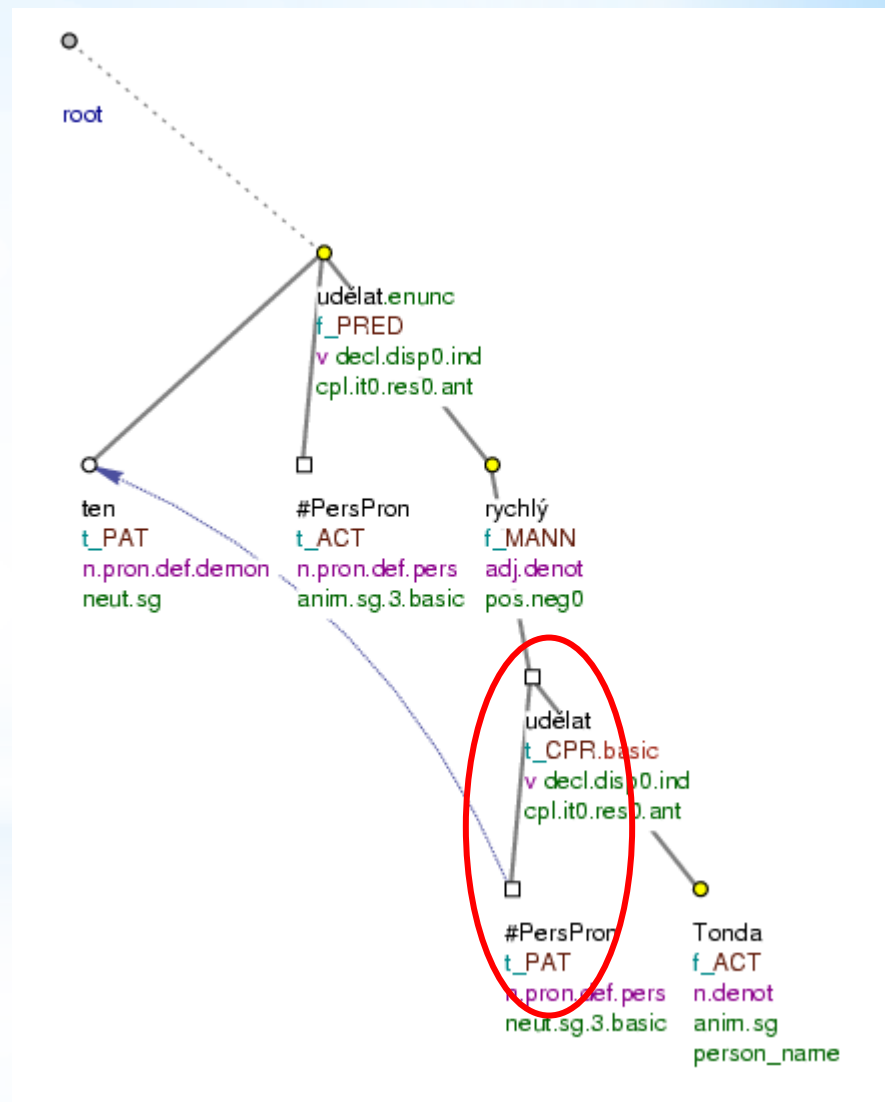
cz. *Klienti pojišťoven, které ukončí svou činnost, se automaticky vrátí k Všeobecné.*

lit. *Clients of insurance companies which shut down will automatically return to the General {one}.*



Восстановлена синтаксическая конструкция

*Udělal to rychle jako Tonda.
(=lit. (He) did it fast like
Tonda)*



* Актуальное членение (TFA)

* Актуальное членение (ТФА)

- * отражает когнитивную стратегию «данное» - «новое», но относится к системе отдельных языков
- * описание АЧ - неделимая часть описания значения предложения (Sgall и др. 1986) = АЧ семантически релевантно
- * в утвердительном предложении Focus сообщает что-то о Topic : F(T)

* Актуальное членение (TFA)

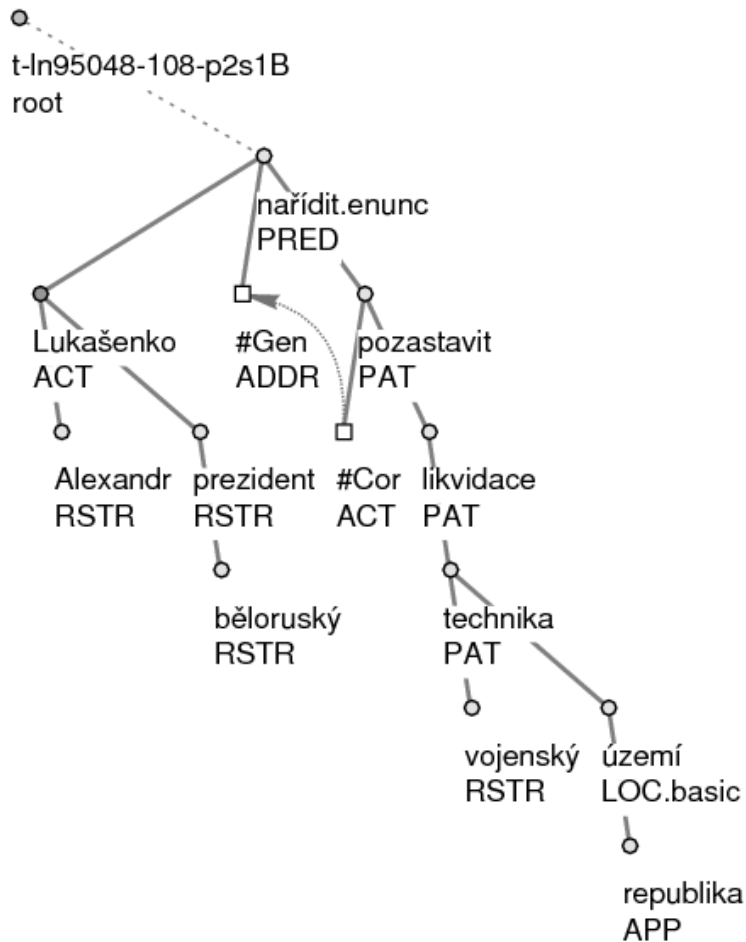
* Контекстная учтенность, активированность в памяти говорящего (атрибут **TFA** на тектограмматическом уровне):

* **t** неконтрастивный контекстно учтенный (активированный в памяти) узел (contextually bound node)

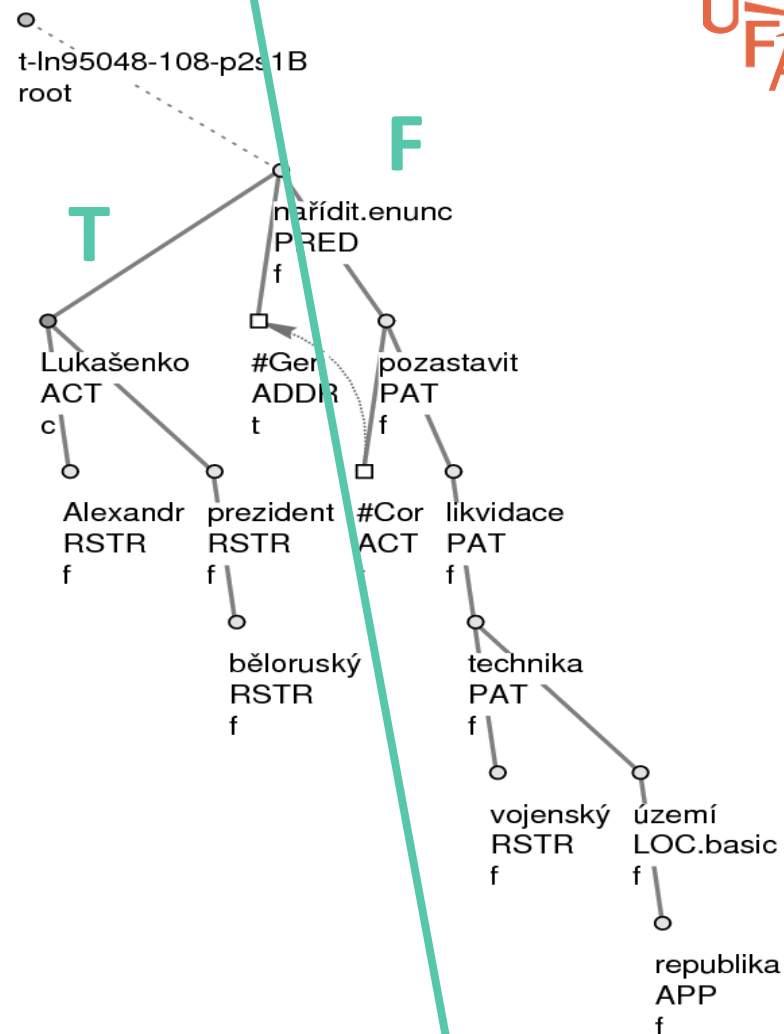
* **c** неконтрастивный контекстно учтенный узел (contrastive contextually bound node)

* **f** контекстно учтенный узел (contextually non-bound node)

* Функциональная перспектива предложения (атрибут **deepord** на тектограмматическом уровне)



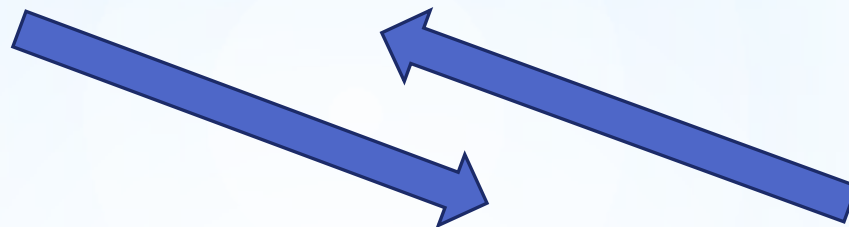
Běloruský prezident Alexandr Lukašenko nařídil pozastavit likvidaci vojenské techniky na území republiky.



Běloruský prezident Alexandr Lukašenko nařídil pozastavit likvidaci vojenské techniky na území republiky.

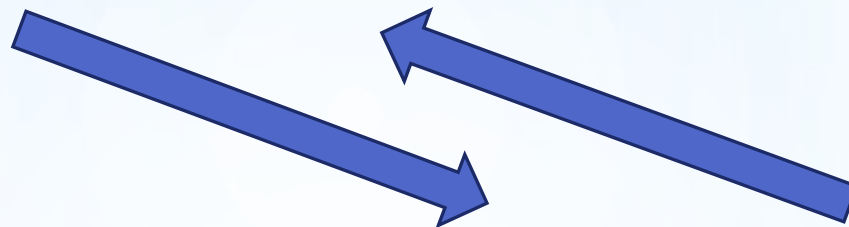
Белорусский президент Александр Лукашенко велел остановить процесс ликвидации военной техники на территории республики.

актуальное членение



коррелятивность

актуальное членение



корелферентность

- (1) контекстно учтенные (contextually bound) выражения без анафорической стрелки
- (2) salience - расчет активированности с учетом корелферентных цепочек и активированности референта в тексте - rule-based

*Contextually bound NP без анафорической стрелки

	В PDT 3.0	%
Всего узлов с tfa = t	69583	100
Узлов с tfa = t без анафорической стрелки	21529	30
Всего узлов с tfa = t с анафорической стрелкой	48054	70
Из них с текстовой кореферентностью	37606	54
Из них с бриджингом	6775	10
Из них с грамматической кореферентностью	3709	5
Из них со ссылкой на сегмент текста	876	1

Contextually bound NP без анафорической стрелки

	В PDT 3.0	%
Всего узлов с tfa = t	69583	100
Узлов с tfa = t без анафорической стрелки	21529	30
Всего узлов с tfa = t с анафорической стрелкой	48054	70
Из них с текстовой кореференцией	37606	54
Из них с бриджингом	6775	10
Из них с грамматической кореференцией	3709	5
Из них со ссылкой на сегмент текста	876	1

* Причины отсутствия анафорической стрелки



- * Отношение есть, но не размечается

Страховые компании не имеют права продавать свои продукты частным лицам.

- * Известность дает ситуация (экзофора, ситуативные уникаты, вытекает из общего знания о мире и др.)

Интервью провел Василий Лебедев.

- * Вторичные обстоятельства (кулисы)

В феврале прошлого года фирма вышла на международный уровень.

- * малореферентные ИГ (проценты, пропорции, сущ. меры и т.д.)

За прошлый год страховые компании продали 100 тонн пшеницы.

- * Другие причины (в основном, технические)

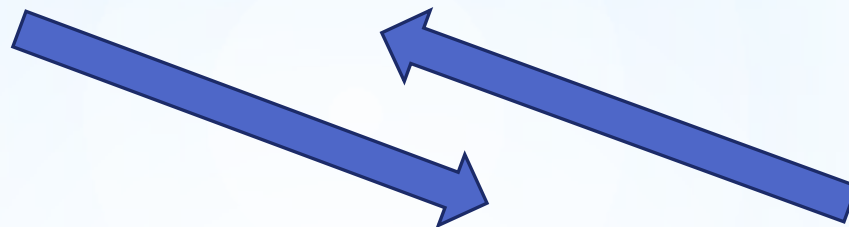
Министерство финансов согласилось с запретом на продажу марихуаны.

* Причины отсутствия анафорической стрелки



	из 500 предл.	%
Отношение есть, но не размечается	157	32
Вторичные обстоятельства (кулисы)	110	22
Известность из ситуации	30	6
Малореферентные ИГ	102	20
Технические причины	88	18
Ошибки	13	2
Всего рассмотрено	500	100

актуальное членение



корелферентность

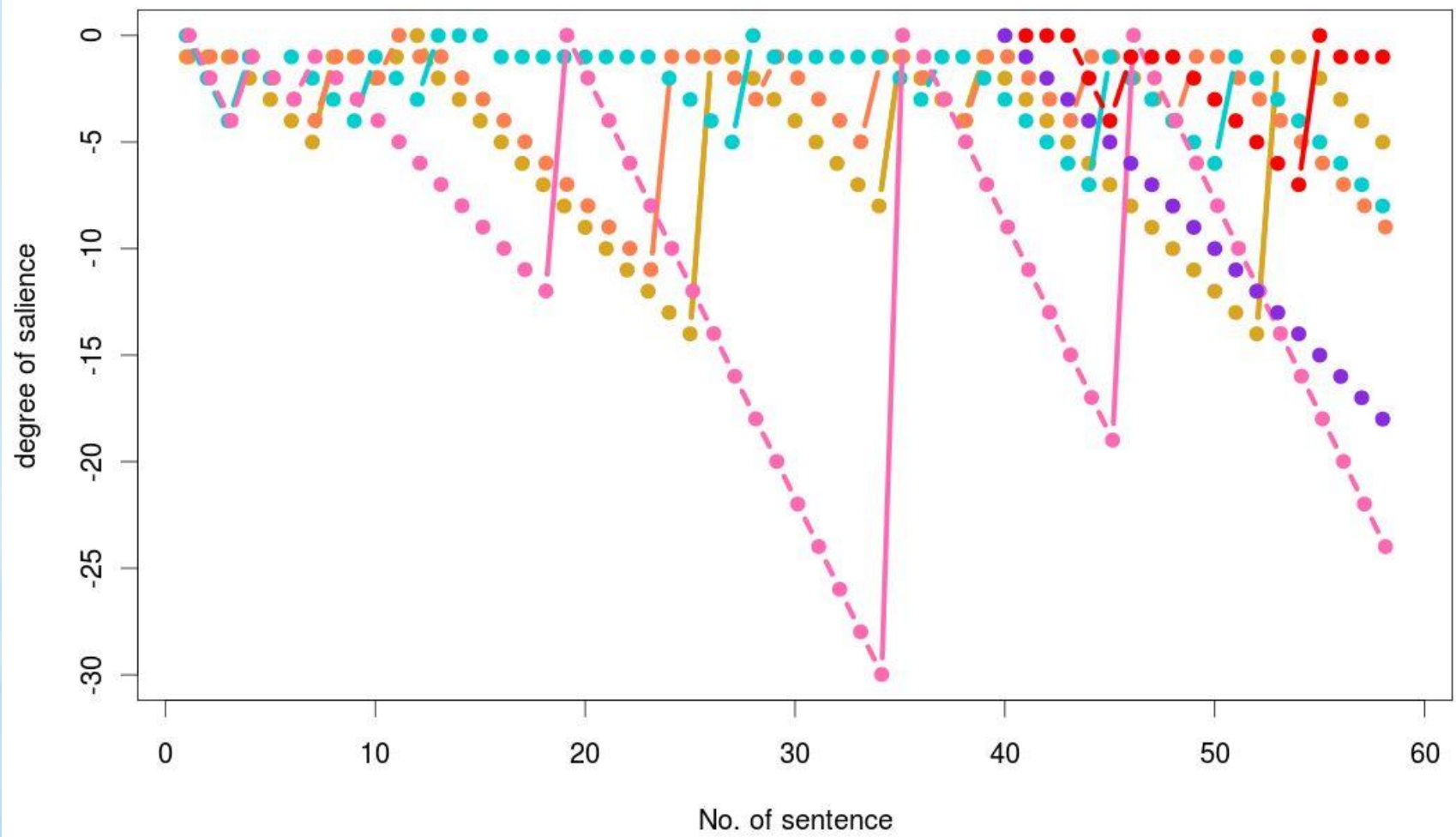
- (1) контекстно учтенные (contextually bound) выражения без анафорической стрелки
- (2) salience - расчет активированности с учетом корелферентных цепочек и активированности референта в тексте - rule-based

* TFA + кореферентность = salience предварительные правила

- (i) if r is expressed by a weak pronoun (or zero, i.e. deleted in the surface shape) in a sentence, it retains its salience degree after this sentence is uttered: $dg_i(r) := dg_{i-1}(r)$;*
- (ii) if r is expressed by a noun (group) carrying nb, then $dg_i(r) = 0$;*
- (iii) if r is expressed by a noun (group) carrying cb, then $dg_i(r) = 1$;*
- (iv) $dg_i(q) := dg_i(r) + 2$ obtains for every referent q that is not itself referred to in S_i , but is immediately associated with an item present here;*
- (v) if r neither is included in S_i , nor refers to an associated object, and has been mentioned in the focus of the preceding sentence, then $dg_i(r) := dg_{i-1}(r) + 2$.*
- (vi) if r neither is included in S_i , nor refers to an associated object, and has been mentioned in the topic of the preceding sentence, then $dg_i(r) := dg_{i-1}(r) + 1$.*

*Salience -Текст

1. Across the river **they** could now see a fire with **two figures** beside it. 2. When they moved closer, 3. they could make out two white horses against the background of the dark bushes. 4. Then he recognized them. 5. The pale blue **buggy**. 6. Two hours ago, **the beauty** from Chicago had sat on the seat. 7. While the **black man** in livery had gone into Kapino's for beer. 8. They stopped ... 9. and looked across the river. 10. The young lady in the white dress was biting into a chicken leg. ... 11. **He** looked at **Magda**. 12. The child's eyes, wide in amazement, stared across the river at this fairytale banquet. ... 13. He looked at the straw hat. 14. Yes, beside it in the grass a pair of white shoes had been casually tossed. 15. and beside them lay a crumpled white pile. ... 31. and ... she had lifted the skirt over her head, 32. slipped out of it 33. and stood there in nothing but white knee-length knickers ... 34. He couldn't take his eyes off her. ... 35. From downstream they could hear a banjo playing. 36. A pleasant baritone voice sang: „...“ 37. The girl let her hands drop... 38. Cautiously, she stepped into the water. 39. On their side of the river, ..., something creaked. 40. Looking towards the sound, he could barely distinguish the outline of a small **rowboat** 41. and, in it, **someone's dark silhouette**. 42. The moonlight fell on the head, the white whiskers, the hair in disarray. 43. The Master! 44. He looked quickly across the stream 45. and saw the Rusalka up to her waist in the water. „Borne like a vapour ...“ 46. The Master's head turned in profile towards the velvet baritone. 47. He doesn't see; 48. he only hears, 49. he thought. 50. He himself saw. ... 51. The Rusalka was slowly lowering herself into the water, ... 52. Finally, all that remained on the water was a burning waterlily. 53. Suddenly the child saw too 54. and shrieked, 55. „Papa!“ 56. The Master started, 57. looked around 58. and then saw.



* Дискурсивная разметка - коннекторы, аргументы, деревья

- * *"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.

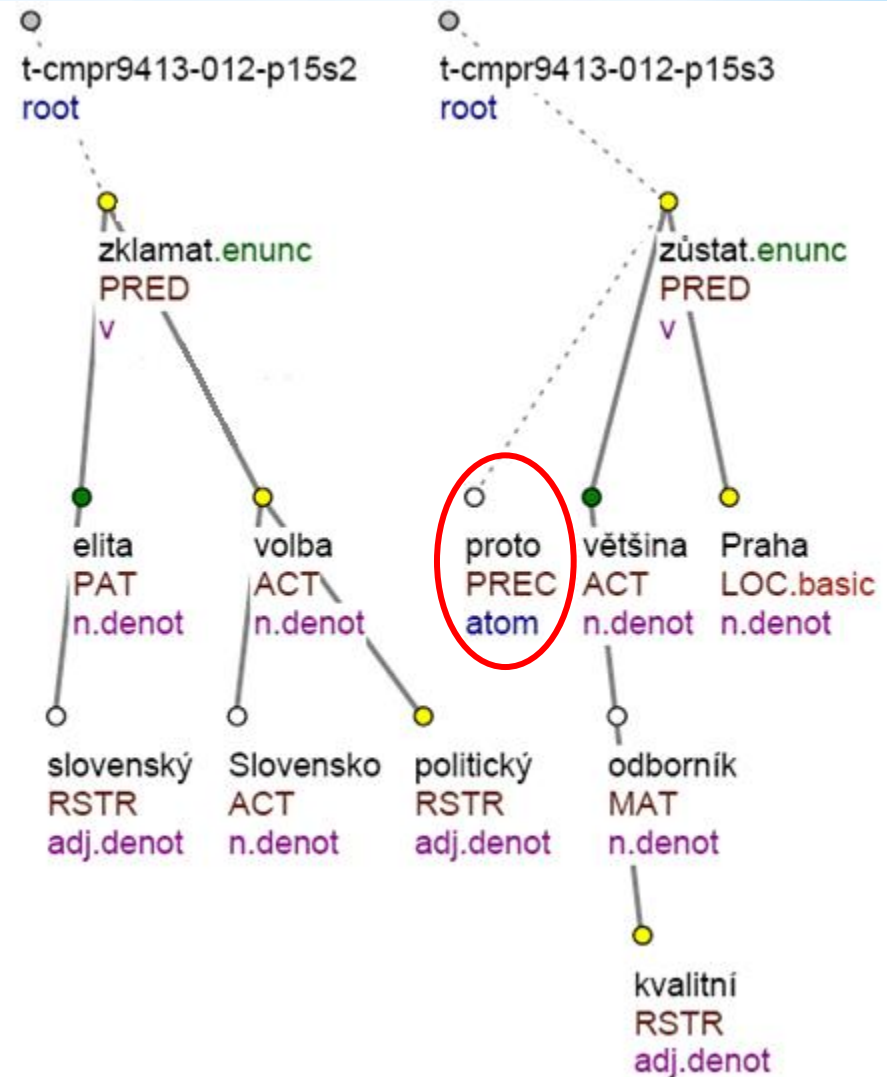
- * ***"Slovak elite was disappointed by the political choice of Slovakia. Therefore, the majority of quality specialists remained in Prague."***

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."

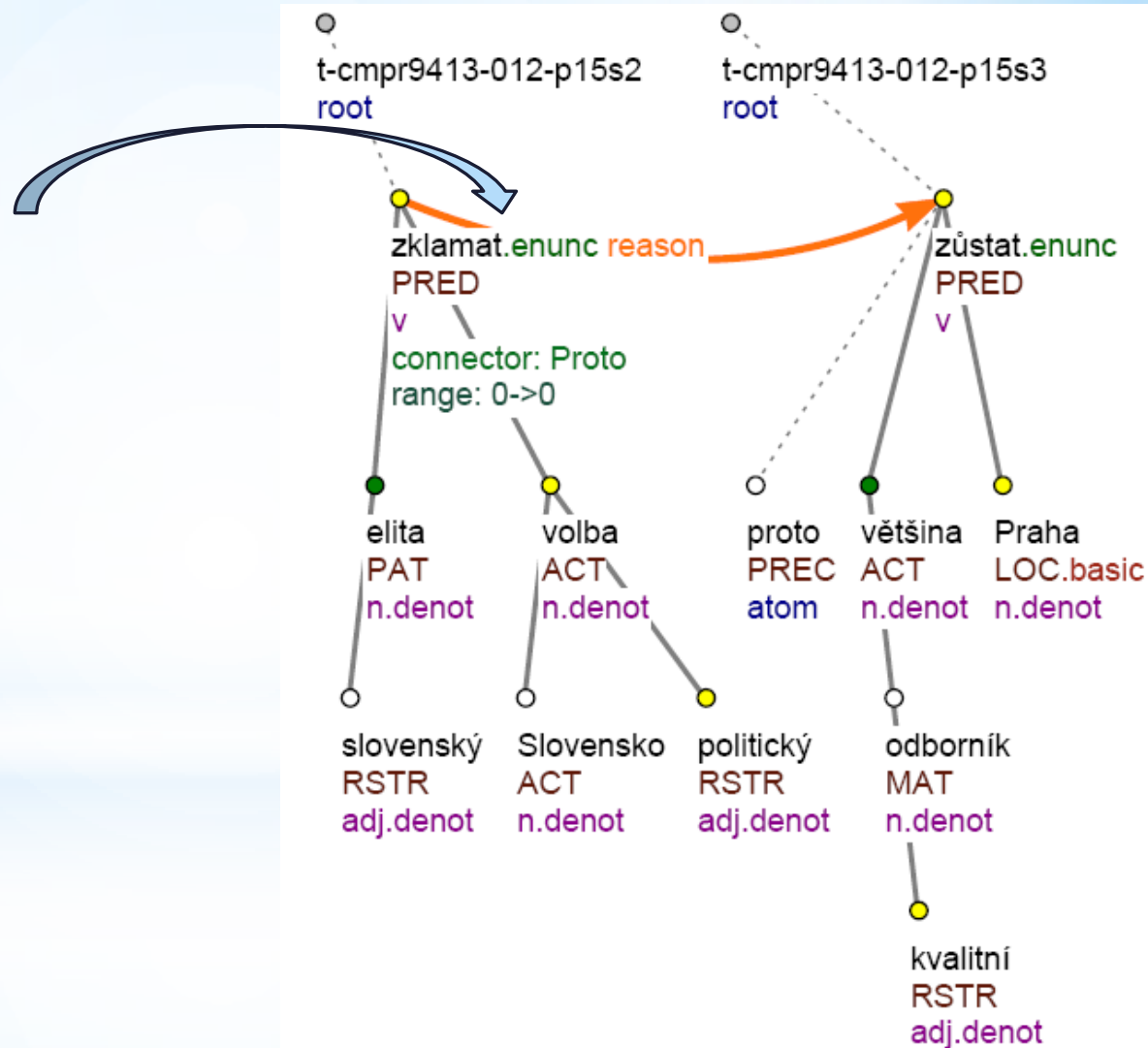
Словацкая элита была разочарована политическим выбором
Словакии. Поэтому большинство хороших специалистов
осталось в Праге.

- * *"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."*

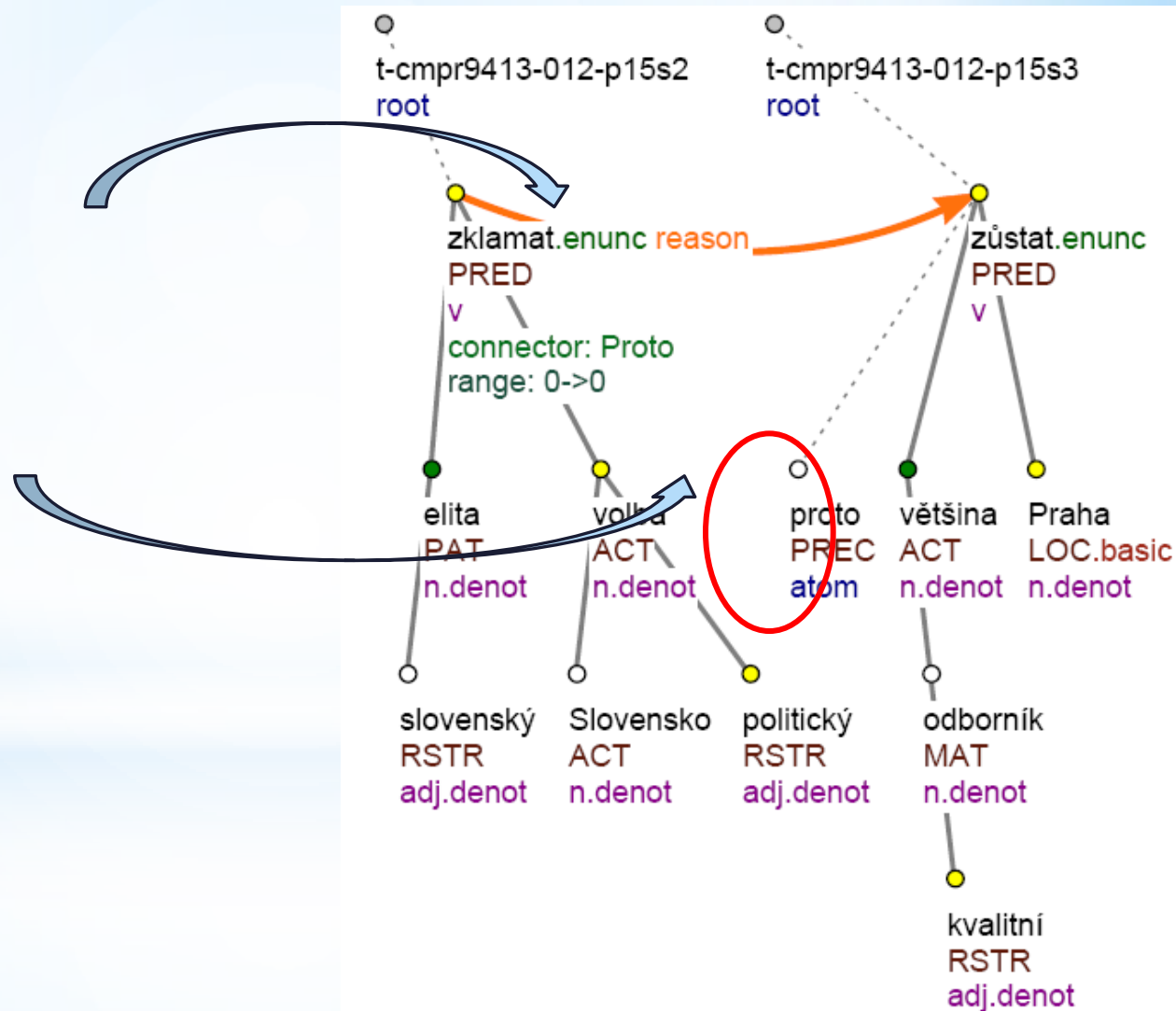
- * аргументы связываются стрелкой



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

* аргументы связываются стрелкой

* добавляем коннектор

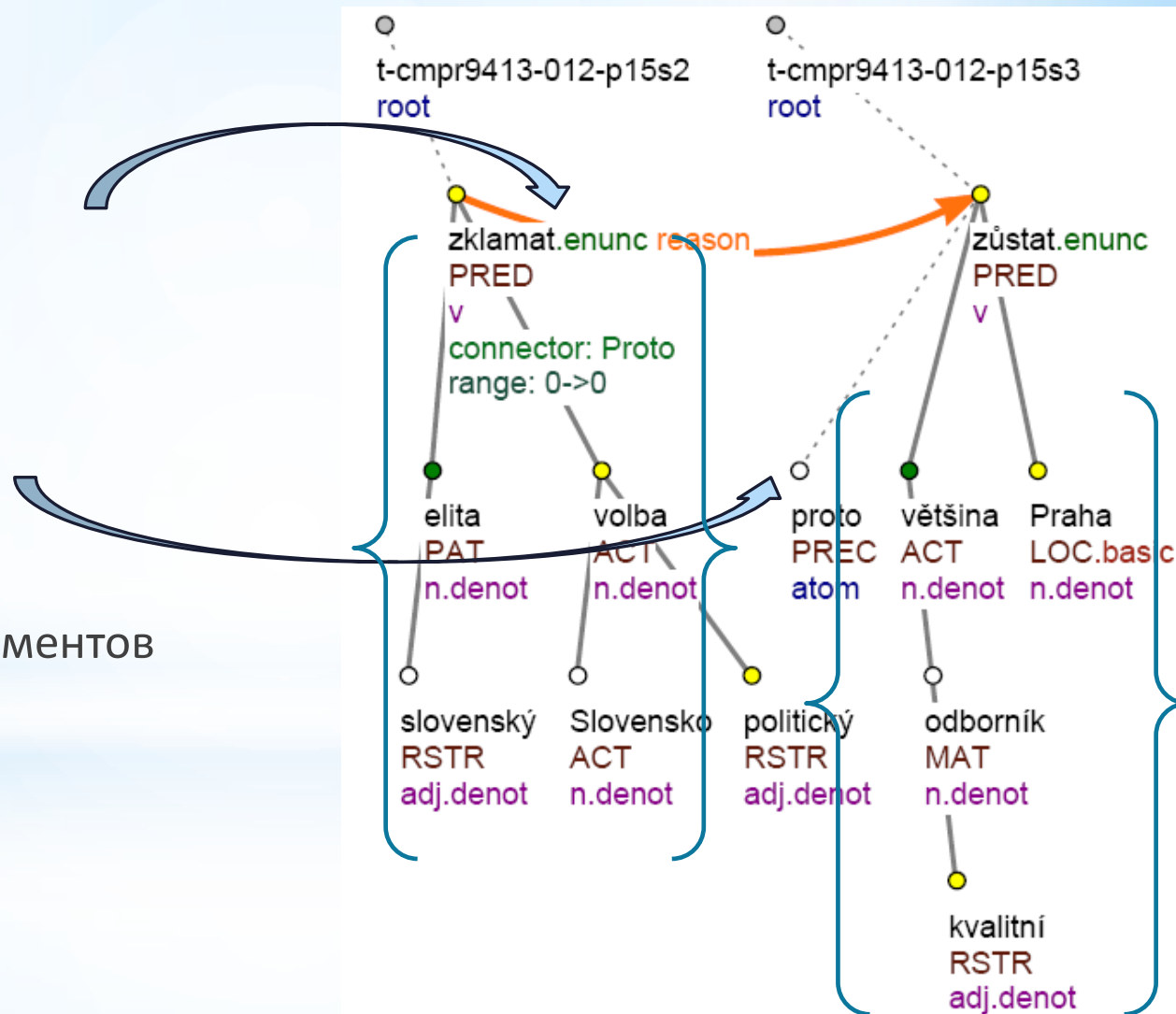


* *"Slovak elite was disappointed by the political choice of Slovakia.
Therefore, the majority of quality specialists remained in Prague."*

* аргументы связываются стрелкой

* добавляем коннектор

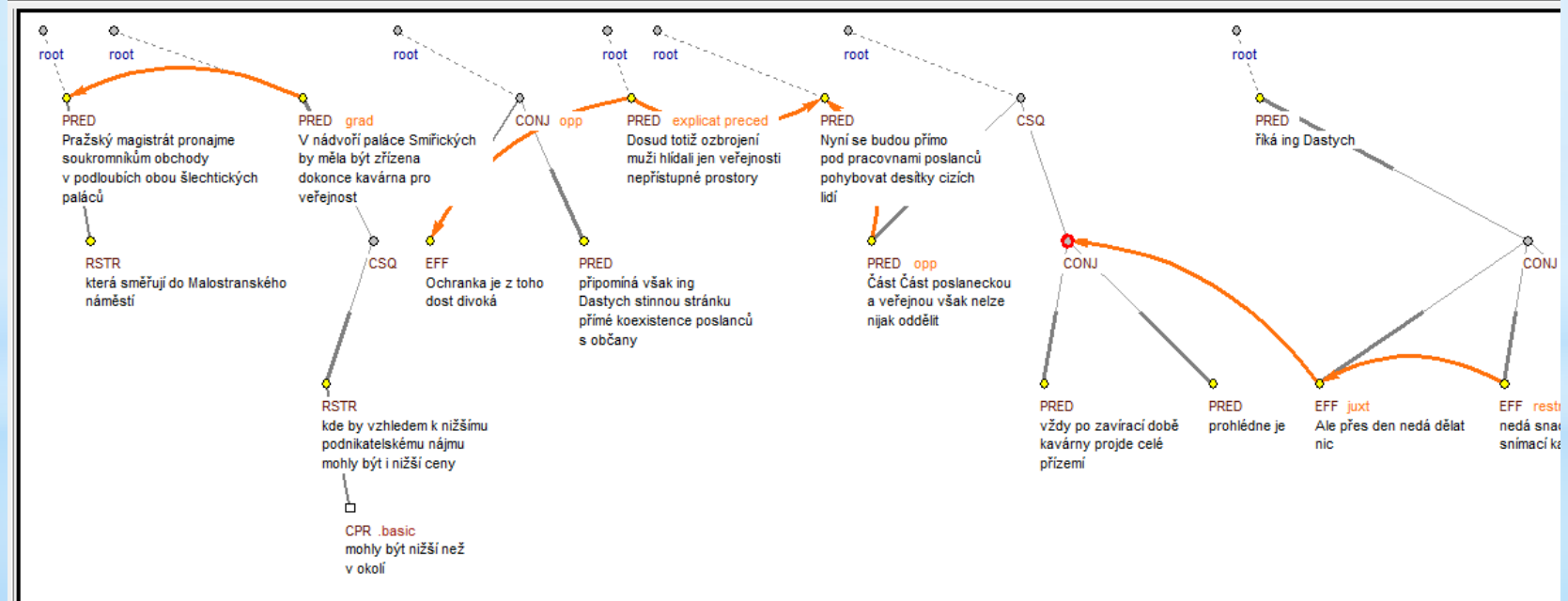
* обозначаем размер аргументов



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.

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

*Примеры



* *Ваня с Машей два года встречались. Год назад завели ребенка. Потом они решили пожениться.*

Temporal → Precedence

* *Я испеку пирог. Но сначала купи яйца.*

Contingency → Condition

* *По вечерам она не сидела дома. Могла, например, пойти в кино.*

Expansion → Instantiation

* *Методика зависит от вас. Просто сделайте так, как вам удобно.*

Expansion → Equivalence

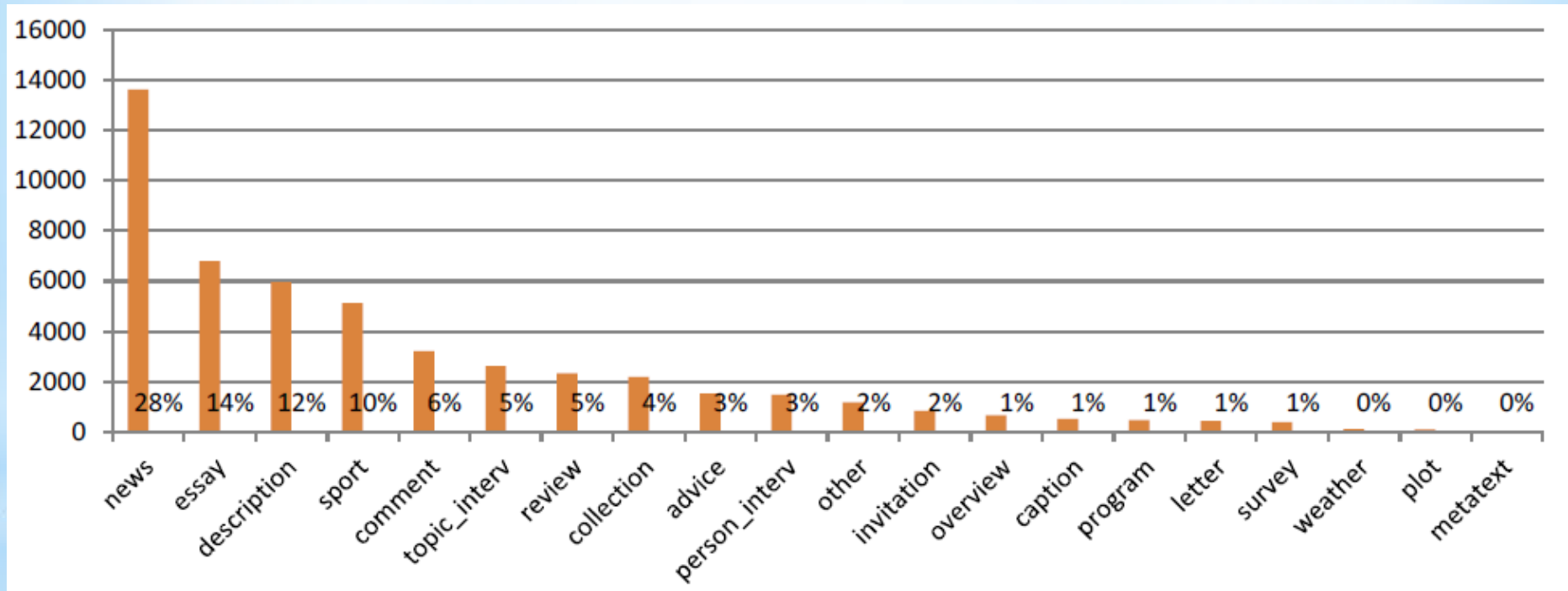
Дискурс, Актуальное членение и кореферентность



- * проблема контраста в АЧ, кореферентности и дискурсе
- * взаимодействие кореферентности и дискурса; многословные дискурсивные маркеры (AltLex)
- * тектограмматика и дискурс, тектограмматика и кореферентность

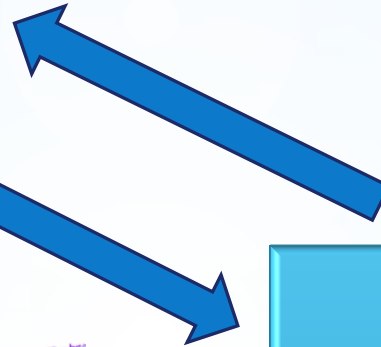
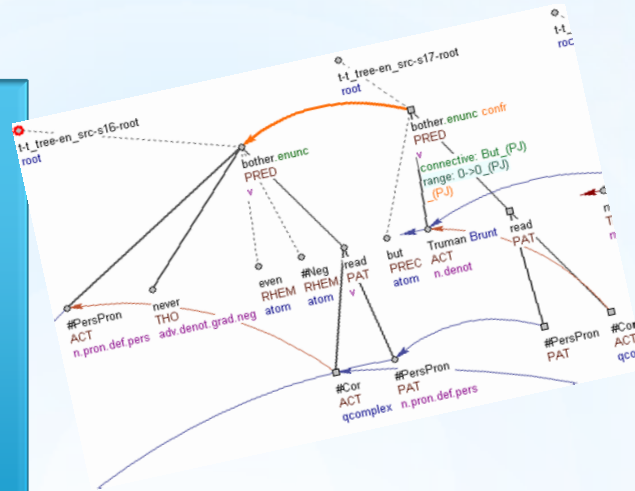
Genre classification

* размечено вручную (Poláková и др. 2014, LREC)



PDT

Prague Dependency Treebank



GECCo

German-English Contrasts in Cohesion

far out on the frozen tip of the continent even the polar bears couldn't find him? Betr
 ve title again with a slow studied movement of his lips. It was only words, only history.
 s. He'd never even bothered to read it. But Truman had. At the moment Truman wa
 air of unconcern, as if showing his alienated son the work of his mad wasted life was ar
 ill gin and lemonade that he was wrought up. The old man suddenly darted a glance ov
 hering revelation, his father the phantom con

reference [conf] ellipsis substitution

type connect

func adversative

problematic ☒ no ☐ yes

☒ Suppress check ☒ Warn on extra attributes

PDT

- базируется на Functional Generative Description (Sgall et al.) and Penn-style discourse annotation (Joshi, Prasad et al.)
- тексты на чешском
- журн. жанр, письменные тексты (са. 50,000 предл.)
- разметка на трех уровнях – морф., пов.-синт., гл.-синт.
- разработана для NLP и лингв. анализа

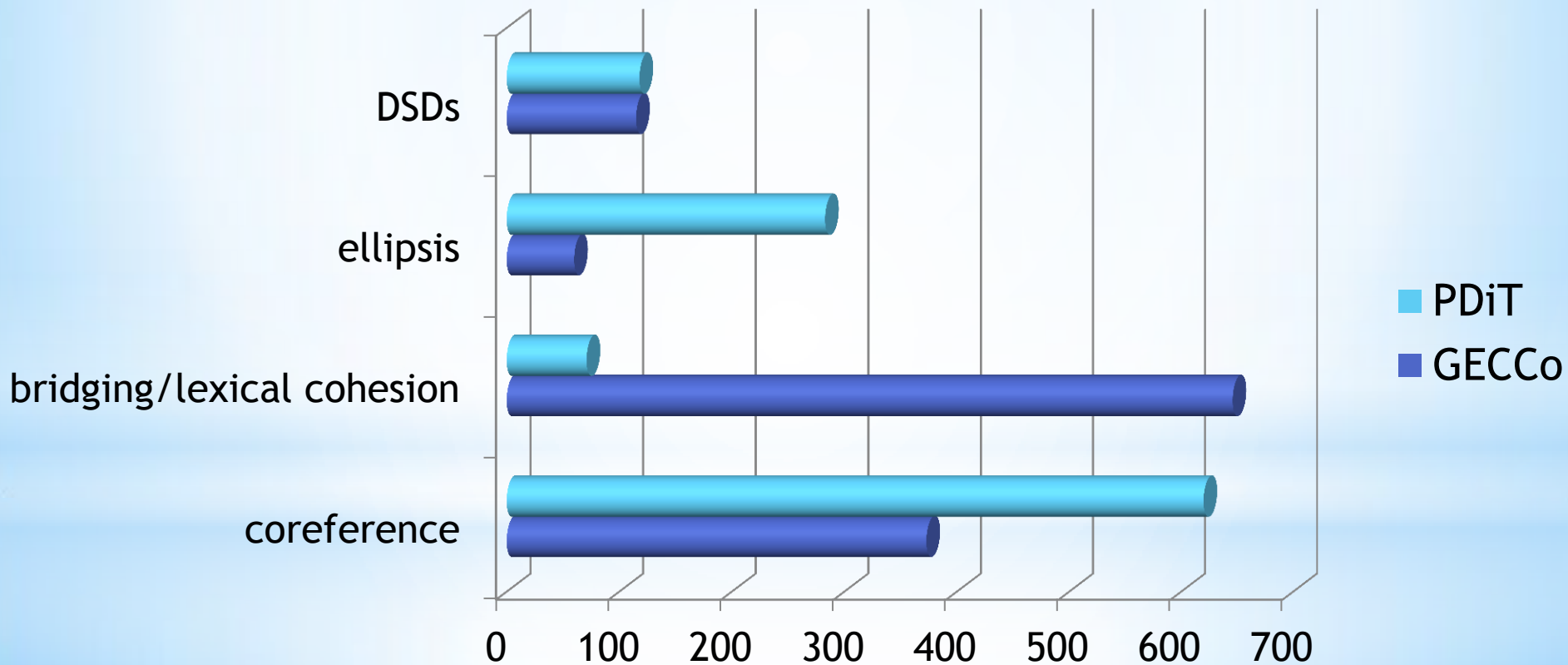
GECCo

- базируется на определении средств когезии по Halliday & Hasan (1976) в англ.
- параллельные и сравнимые тексты в англ. и нем.
- разные регистры и жанры, вкл. разговорные и письменные тексты (ок. 80,000 предложений)
- морфосинтаксическая разметка
- разработан для сравнительного анализа англ. и нем. языков

PDT vs. GECCo предмет разметки

GECCo	(co)reference	lexical cohesion	substitution	ellipsis	conjunctive relations
PDiT	coreference	bridging	-	ellipsis captured in dependency trees	connectives, arguments, relations

* Статистика отношений



precedence - succession	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	CONTINGENCY	COMPARISON (CONTRAST)	EXPANSION

discourse markers
(attitude markers,
modal particles)

TEMPORAL	CAUSAL	ADVERSATIVE	ADDITIVE	MODAL
temporal relation between events	relation of causality/dependence between	relation of contrast/alternative, for two events which are not true at the same time	relation of addition, for two events that are true/not true at the same time	relation between events connected by an evaluation of the speaker
<i>after, afterwards, at the same time..</i>	<i>because, therefore, that's why...</i>	<i>yet, although, by contrast...</i>	<i>and, in addition...</i>	<i>well, sure, of course, surely, eventually...</i>
<i>nachdem, danach, gleichzeitig..</i>	<i>weil, deshalb, aus diesem Grund..</i>	<i>doch, obwohl, im Gegensatz dazu..</i>	<i>und, außerdem..</i>	<i>klar, sicher, allerdings, jedenfalls, eigentlich, wohl...</i>

* Выводы:



- * много всего размечено
- * все похоже, схемы сравнимы
- * есть, над чем работать
- * интересно делать что-то с русским

Спасибо за внимание

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