

Towards comprehensive approaches to discourse-structuring devices

PDiT - GECCo Scientific mission

Anna Nedoluzhko, Ekaterina Lapshinova, Kerstin Kunz

Background Information

- STSM from Prague to Saarbrücken, January 16 - February 8, 2015, within the COST Action ‘Textlink - Structuring Discourse in Multilingual Europe’, a EU initiative

The aims of STSM:

- identify commonalities and differences between the two frameworks
  - Prague Discourse Treebank (PDiT)
  - GECCo (German-English Contrasts in Cohesion) at Saarland University

- to explore the interoperability of our approaches, to find benefits and drawbacks of each approach and the ways to improve them

General Description of Approaches

PDiT

- based on Functional Generative Description (Sgall et al.) and Penn-style discourse annotation (Joshi, Prasad et al.)
- texts in Czech

- journalistic written texts with further genre classification (ca. 50,000 sentences)
- multilayer annotation: morphological, analytical and tectogrammatical
- elaborated for various NLP-tasks and linguistic analysis

GECCO

- based on the definition of cohesion and cohesive devices in English by (Halliday & Hasan, 1976)
- comparable and parallel texts in English and German
- various registers, including written and spoken dimensions (ca. 80,000 sentences)
- complex morpho-syntactic annotation

- elaborated for a contrastive analysis of two languages

General Description of Schemes - PDiT

- discourse annotation (explicit connectives + arguments, sense tags (= PDTB)
- coreference annotation (pronominal coreference, NP-coreference, event-anaphora, zero anaphora)
- bridging relations
- Information Structure (topic - focus articulation)
- ellipsis

General Description of Schemes - GECCo

- **Cohesive relations** require a linguistic trigger, a **cohesive device** which explicitly signals that there is a relation to another textual expression
- **Cohesive devices** can be 1) grammar- or 2) vocabulary-driven:

1) semantic reduction of expressions to functional items which are syntactically obligatory
2) lexical vocabulary of the discourse segment

- **Cohesive devices**: conjunctive relations, reference, substitution, ellipsis and lexical cohesion, as well as their structural, functional subtypes and further properties
- **Cohesive relations**: coreference chains, lexical chains, and also links between elliptical expressions and their antecedents

GECCo <--> PDiT Concepts

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

Experiment Settings

Data (200 sentences):

- 1 longer fictional text from the GECCo corpus (EO_FICTION_004)
- 4 shorter journalistic texts from PCEDT (wsj_0022, 0039, 0088, 0094)

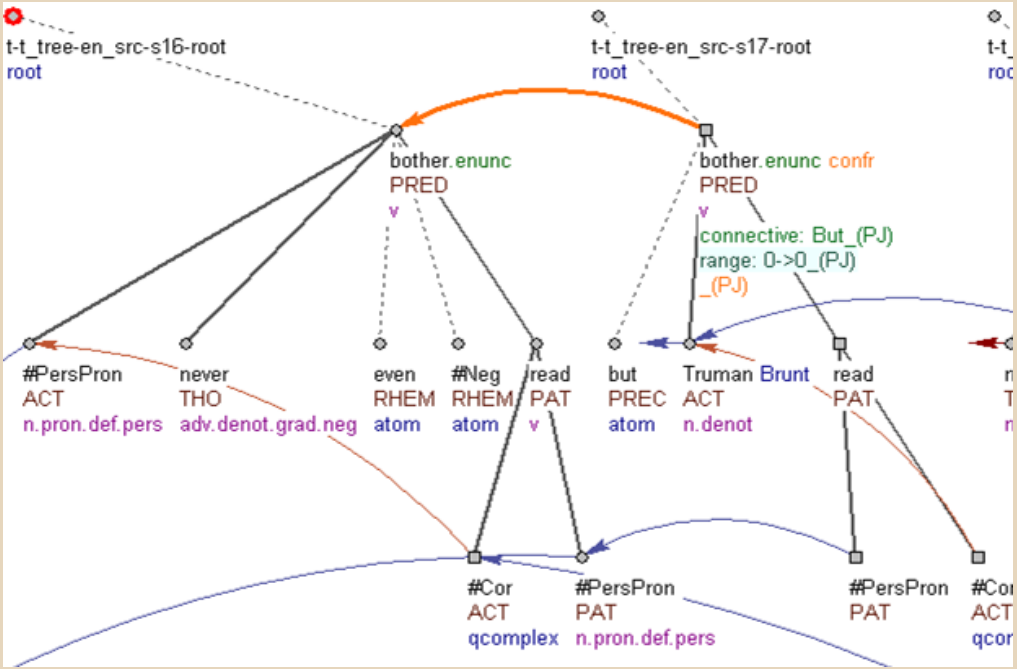
Tools:

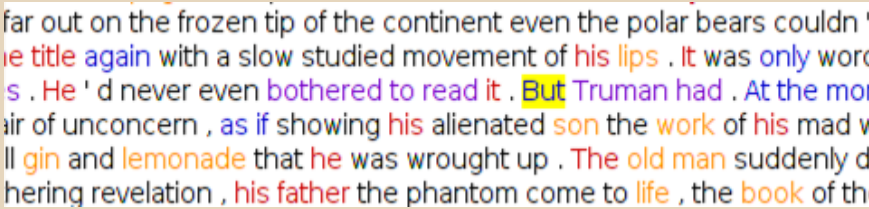
- MMAX2 (GECCo)
- Tred (Prague)

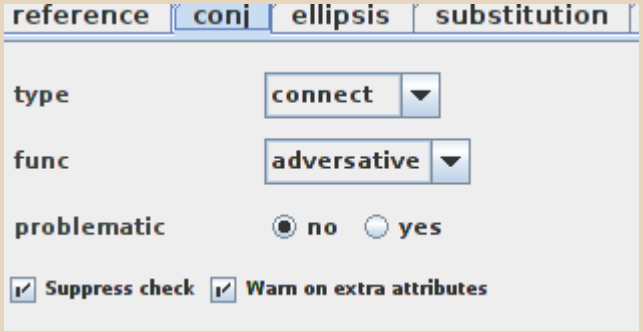
Annotation:

- manual
- 4 annotators in GECCo,
- 4 annotators in PDiT

Visualization in Tools





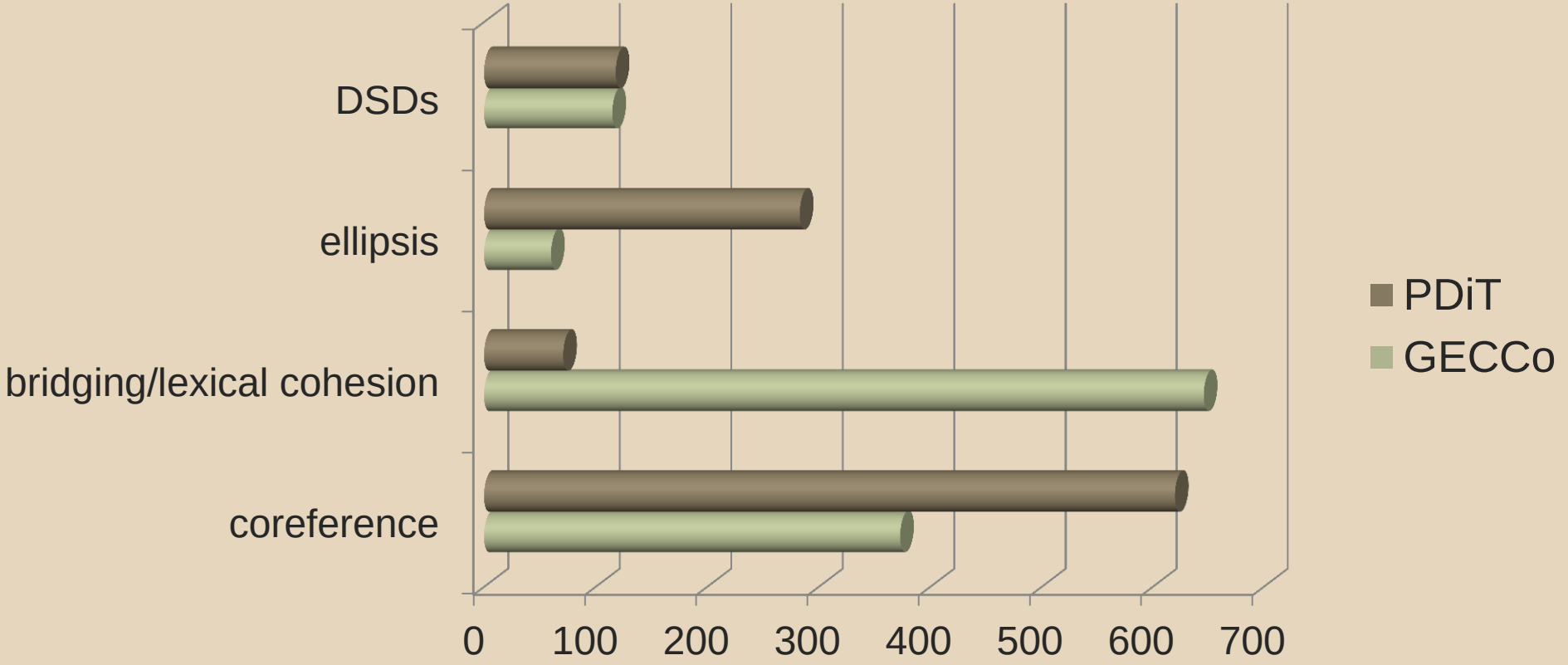


He 'd never even bothered to read it. ***But*** *Truman had.*

Overall Statistics

	GECCo		PDiT	
	wsj texts	fiction	wsj texts	fiction
coreferring expressions	188	185	245	247
bridging/lexical cohesion	417	229	25	46
substitution	2	3	-	-
ellipsis	13	47	142	141
DSDs	60	55	68	48

Overall Statistics



Overall Statistics

Discourse-Structuring Devices

	GECCo	PDiT
framework behind	SFL, grammars	PDTB
marking arguments	no	yes
explicit / implicit	only explicit	
semantic labels on	connectives	both arguments
set of connectives	closed/open	open (vs. PDTB)
alternative lexicalizations	captured by other cohesive devices	yes

Discourse Annotation - PDiT

- lexically-grounded approach of identification of discourse connectives, discourse units (propositions) linked by them and semantic relations between these units
- Penn Discourse Treebank (PDTB, Prasad et. al., 2008) - "shallow discourse parsing“, identification of discourse markers and relations they express

Semantic Labels - PDiT

TEMPORAL	CONTINGENCY	COMPARISON (CONTRAST)	EXPANSION
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

Conjunctive Relations in GECCo

- Conjunction concerns the **logico-semantic relations** between propositions, e.g. addition, contrast, cause, etc.
- Definition and classification are based on Halliday and Hasan (1976)
- Categories existing in both English and German, e.g.:

❖ *GO: Sie wollen ein starkes Europa in der Welt. Deshalb hat Großbritannien eine europäische Sicherheitspolitik mit auf den Weg gebracht.*

❖ *EO: They want Europe to be strong in the world. That's why Britain has helped launch a European security policy.*

Semantic Labels in GECCo

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

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) -
in PDiT not considered as connectives

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
after, afterwards, at the same time..	*because, therefore, that's why...*	*yet, although, by contrast...*	*and, in addition...*	*well, sure, of course, surely, eventually...*
nachdem, danach, gleichzeitig..	*weil, deshalb, aus diesem Grund..*	*doch, obwohl, im Gegensatz dazu..*	*und, außerdem..*	*klar, sicher, allerdings, jedenfalls, eigentlich, wohl...*

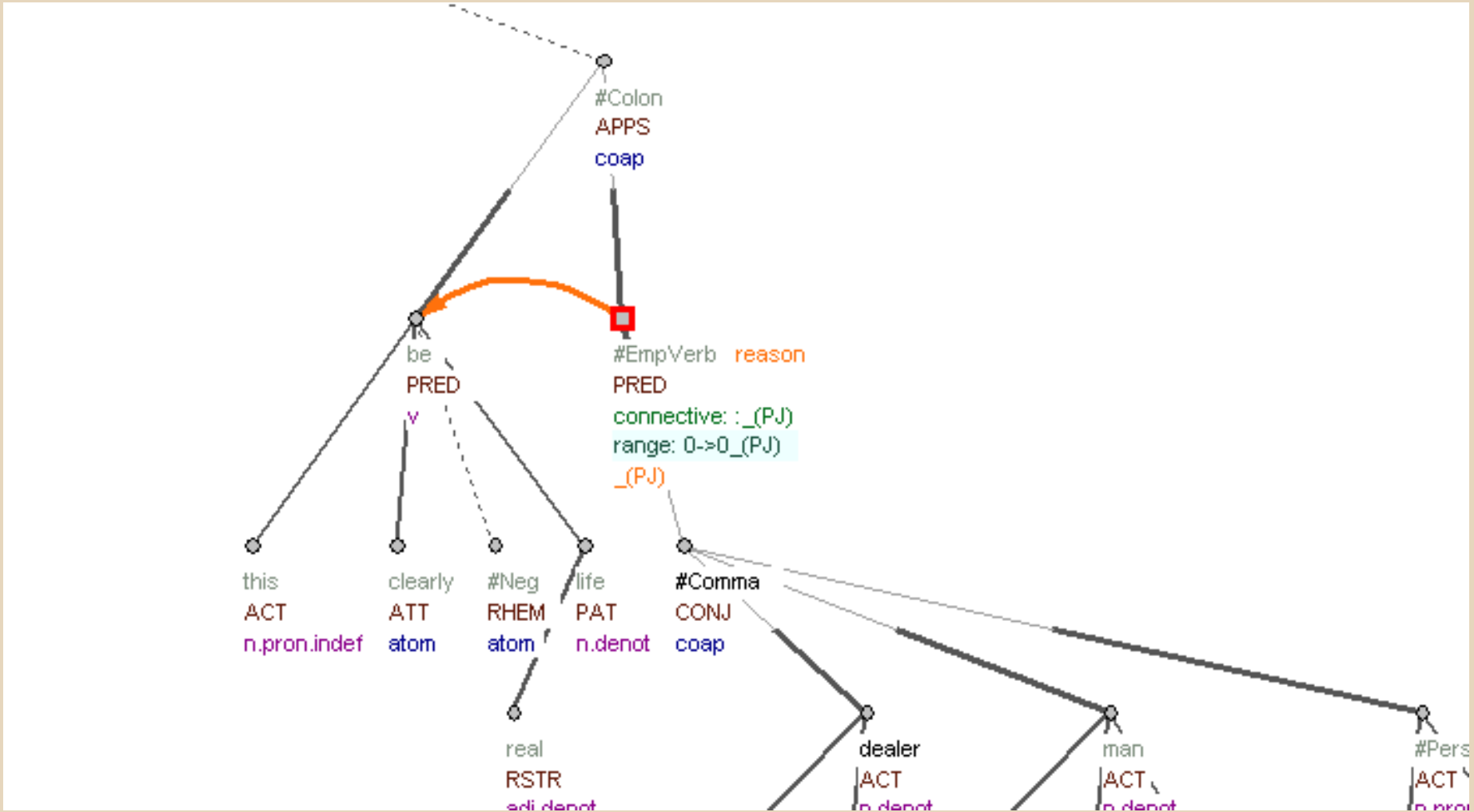
Statistics for DSDs

	GECCo		PDiT	
	wsj texts	fiction	wsj texts	fiction
TEMPORAL	6	11	5	5
CONTINGENCY / CAUSAL	9	6	19	4
COMPARISON (CONTRAST) / ADVERSATIVE	16	10	15	17
EXPANSION / ADDITIVE	22	24	19	22
MODAL	7	4	*not annotated as connectives*	

Examples – DSD vs. non-DSD

This clearly is not real life: no crack dealers, no dead-eyed men selling four-year-old copies of Cosmopolitan, no one curled up in a cardboard box

PDiT: reason-result, ellipsis
GECCo: ellipsis


#Colon
APPS
coap
be
PRED
v
#EmpVerb reason
PRED
connective: :_(PJ)
range: 0->0_(PJ)
_(PJ)
this
ACT
n.pron.indef
clearly
ATT
atom
#Neg
RHEM
atom
life
PAT
n.denot
#Comma
CONJ
coap
real
RSTR
dealer
ACT
man
ACT
#Pers
ACT

Examples – Different Types

William Gates and Paul Allen in 1975 developed an early language-housekeeper system for PCs, and Gates became an industry billionaire six years after IBM adapted one of these versions in 1981.

PDiT: *reason-result*
GECCo: *additive relation*

Results

- Categories annotated in both approaches seem to partly depend on the **text/genres/registers**
- **the greatest difference lies in** lexical cohesion and coreference
- **Different conceptions** are reflected in the annotation

Results and Reasons

- conceptions for two distant languages, differences in information structure in English and Czech: interplay between determination, syntactic constraints and information structure
- all the levels are inter-dependent (differences in numbers for certain categories)

Conclusion and Outlook

- the discovered overlaps provide the possibilities for the creation of an interoperable scheme **applicable across languages and genres**
- our **future** work!
- more information: **WG2 poster session**:
  Tuesday 9.30-10.45, **SOCRATE 11 patio**

THANK YOU!
DĚKUJEME!
DANKE!