Finding Nexus in the PDiT and GECCo Annotation Schemes

Ekaterina Lapshinova-Koltunski*, Anna Nedoluzhko, Kerstin Kunz***, Lucie Poláková**, Jiří Mírovský**, Pavlína Jínová****

Saarland University*, Charles University in Prague**, University of Heidelberg***

e.lapshinova@mx.uni-saarland.de, nedoluzko@ufal.mff.cuni.cz, kerstin.kunz@iued.uni-heidelberg.de,
polakova,mirovsky,jinova@ufal.mff.cuni.cz

BACKGROUND INFORMATION

Aims and Motivation

► compare two frameworks for the analysis and annotation of discourse-structuring devices (DSDs) and further discourse phenomena in
✓ GECCo ✓ PDiT
► identify commonalities and/or differences between the two frameworks

Overarching Goal

► achieve interoperability and creating an 'all-in-one' scheme applicable to different languages, different genres and registers, including spoken and written dimensions
► for the time being: English texts only (sake of convenience)
► for the future: German and Czech (differences between Germanic and Slavic languages)

PDiT

► Functional Generative Description (Sgall et al., 1986) and Penn-style discourse annotation (Prasad et al., 2007)
► journalistic texts (written) in Czech with further genre classification (ca. 50,000 sentences)
► multilayer information:
morphological, analytical and tectogrammatical

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

GECCo

► based on the definition of cohesion and cohesive devices in English by (Halliday & Hasan, 1976) elaborated for a **contrastive analysis** of two languages and different registers (genres)
► comparable and parallel texts in English and German from various registers (written and spoken)
► multilayer information: morpho-syntax

► **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

METHODS AND DATA

Data Description

double annotation:

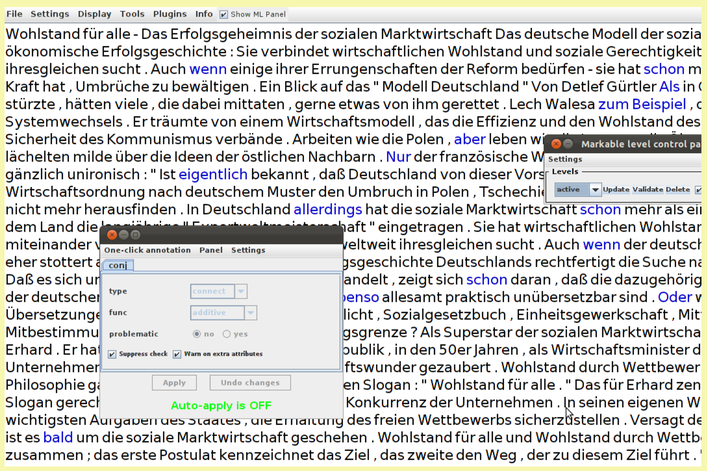
– PDiT scheme (Poláková et al., 2013)
– GECCo scheme (Lapshinova & Kunz, 2014a,b)

the same datasets:

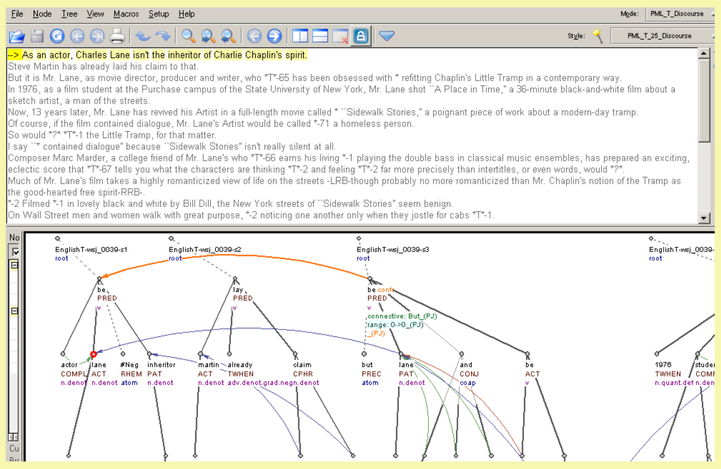
journalistic:
4 shorter texts from PCEDT:
wsj_0022, wsj_0039, wsj_0088, wsj_0094)

fictional:
1 longer text from the GECCo corpus
EO_FICTION_004

MMAX2 (Müller & Strube, 2006)



TrEd (Pajas & Štěpánek, 2008)



GENERAL COMPARISON

Phenomena in Focus

GECCo	(co)reference	lex. cohesion	substitution	ellipsis	**conjunctive relations**
PDiT	coreference	bridging	–	ellipsis in dep. trees	**connectives arguments relations**

Annotation Statistics

	genre	coref.expr.	bridg./lex.coh.	subst.	ellip.	DSD
GECCo	**journalistic**	188	417	2	13	60
	fictional	185	229	3	47	55
PDiT	**journalistic**	317	25	-	142	68
	fictional	303	46	-	141	48

Summary

► **Different conceptions** are reflected in the annotation:
EXAMPLE: annotation of coreferring expressions with modifiers on the basis of explicit signals (e.g by a possessive, a definite article) in GECCo
vs. orientation only on referential identity in PDiT, e.g. *she - her children* (corefer in GECCo, but not in PDiT)
► **Categories** annotated in our two approaches seem to depend on the **genres or registers**, and maybe texts themselves
► **the greatest difference:** lexical cohesion and coreference

► **Reasons:**
⇐ GECCo: no named entities in coref.;
⇐ lexical coh. is mostly based on semantic relation;
⇐ PDiT sometimes includes pragmatic relations
⇐ all the levels are inter-dependent (differences in numbers for certain categories)
⇐ conceptions for two distant languages with no common heritage
⇐ differences in information structure in EN and CZ: interplay between determination, syntactic constraints and information structure

CASE STUDY: DISCOURSE RELATIONS

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 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...*

Conjunctive Relations/Connectives

	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 lexicalisations	other coh.devices	yes

Statistics for Discourse

	GECCo		PDiT	
	journalistic	**fiction**	**journalistic**	**fiction**
temporal	6	11	5	5
contig./causal	9	6	19	4
compar./advers.	16	10	15	17
expans./ additive	22	24	19	22
modal	7	4	not annotated	

Structuring Discourse in Multilingual Europe, 1st action conference, *26.-28. January 2015, Louvain-la-Neuve*