Finding Nexus in the PDiT and GECCo Annotation Schemes

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,mironovskj,ginova@ufal.mff.cuni.cz

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)

METHODS AND DATA

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

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)
- multi-layer 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

Phenomena in Focus

GECCo
- (co)reference
- lex. cohesion
- substitution
- ellipsis
- conjunctive relations

PDiT
- coreference
- bridging
- ellipsis in dep. trees
- connectives
- arguments
- relations

Methods

double annotation:
- PDT 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 ED_FICTION_004

Data Description

MIMAX2 (Müller & Strube, 2006)

TrEd (Pajas & Štěpánek, 2008)

GECCo
- fictional
- journalistic

PDiT
- fictional
- journalistic

Categories

genre coref.expr. bridg./lex.coh. subst. ellip. DSD
GECCo journalistic 188 417 2 13 60
PDiT journalistic 185 223 3 47 55
fictional 317 25 1 142 45
fictional 300 46 - 141 45

GENERAL COMPARISON

Different conceptions are reflected in the annotation:

EXAMPLE: annotation of coreferring expressions with modifiers on the basis of explicit signals (e.g. 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.;
- 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 and discourse markers

Conjunctive Relations/Connectives

- discourse markers (attribute markers, modal particles) in PDiT not considered connectives

Statistics for Discourse

<table>
<thead>
<tr>
<th>Statistics for Discourse</th>
<th>GECCo</th>
<th>PDiT</th>
</tr>
</thead>
<tbody>
<tr>
<td>temporal</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>contig.causal</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>compar./advers.</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>expans./ additive</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>modal</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>