

# **Introduction to Corpus Resources, Annotation and Access: *More Levels of Corpus Annotation***

---

Sabine Schulte im Walde  
Universität Stuttgart

*Foundational Course*  
Departament de Traducció i Filologia  
Universitat Pompeu Fabra  
April 16-20, 2007

# Overview

---

1. Levels of annotation
2. Topic-Focus Articulation
3. Rhetorical structures and discourse connectives
4. Anaphora and coreference
5. Prosodic structure
6. Standardisation projects: MATE and NITE

# Levels of Annotation

---

- Part-of-speech tags
- Lemmata
- Syntactic functions
- Senses
- Semantic roles
- Prosody
- Topic / Focus
- Coreference
- Named Entities
- Discourse relations
- Time
- Emotions ...

# **Topic-Focus Articulation in the Prague Dependency Treebank**

---

# Prague Dependency Treebank (PDT)

---

- Three-level annotation scenario:
  1. **morphological level**
  2. syntactic annotation at the **analytical level**
  3. linguistic meaning at the **tectogrammatical level**
- Crossing the sentence boundary: The human annotators assign (disambiguated) structures, according to the meaning of the sentence in its environment, taking contextual (and factual) information into account.
  - » topic-focus articulation
  - » coreference
- Purpose: linguistic research beyond the sentence limit

# Topic-Focus Articulation (TFA)

---

- TFA reflects the communicative function of the sentence.
- **Topic:** What is the sentence about?  
→ a non-contrastive contextually bound node
- **Focus:** What information about the topic is asserted?  
→ a contextually non-bound node, new information
- An important role is played by the position of the intonation marker.

# Topic vs. Focus

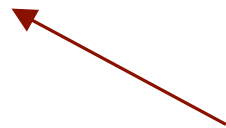
---

- The **topic** (or **theme**) is the part of the proposition that is being talked about (predicated). Once stated, the topic is therefore "old news", i. e. the things already mentioned and understood. The topic is also called theme, and the predicate that gives information on the topic is also called **rheme**.
- The **focus** determines which part of the sentence contributes the most important information. The focus may be highlighted either prosodically or syntactically or both, depending on the language.

# TFA Examples (1)

---

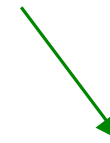
English is spoken in the SHETLANDS.



**topic**



**focus**

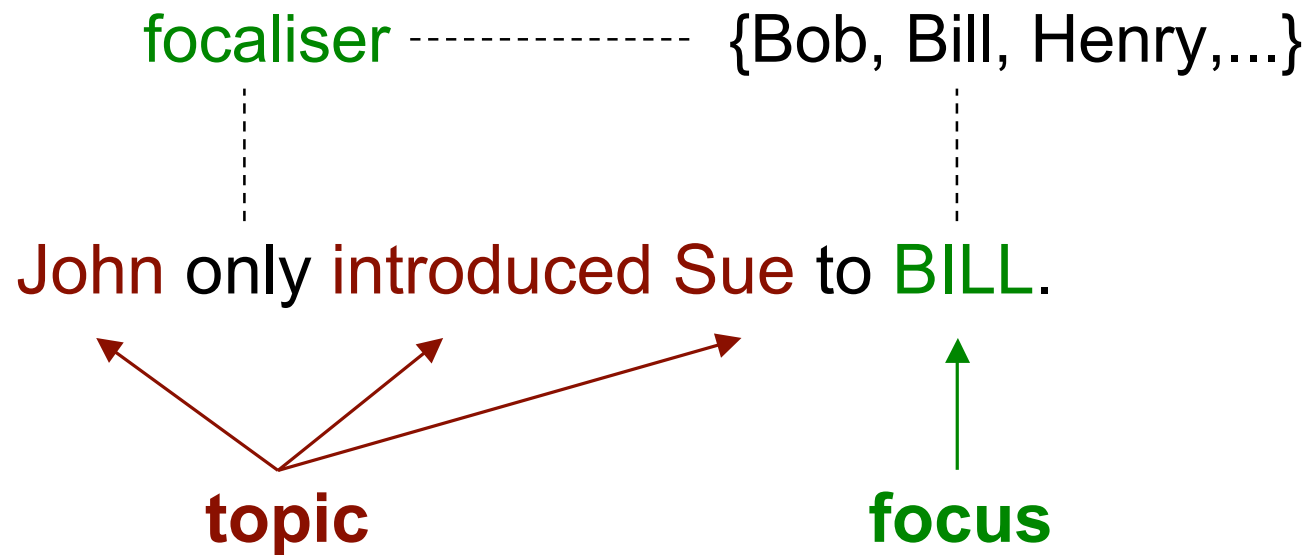


In the Shetlands, one speaks ENGLISH.

(The capitals denote the intonation centre.)

# TFA Examples (2)

---



(The capitals denote the intonation centre.)

# PDT: References

---

- Eva Hajičová (1999): "The Prague Dependency Treebank: Crossing the sentence boundary". In *Proceedings of the 2nd Workshop on Text, Speech, Dialogue*, pp. 20-27. Mariánské Lázně, Czech Republic.
- Eva Hajičová, Jarmila Panevová, and Petr Sgall (2000): "Coreference in annotating a large corpus". In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pp. 497-500. Athens, Greece.
- Oana Postolache, Ivana Kruijff-Korbayová, and Geert-Jan Kruijff (2005): "Data-driven approaches for information structure identification". In *Proceedings of the joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 9-16. Vancouver, Canada.
- PDT online: <http://ufal.mff.cuni.cz/pdt/>

# **Rhetorical Structure: Theory and Corpus**

---

# Rhetorical Structure Theory (RST)

---

- RST is a **theory of discourse structure**.
- RST offers an explanation of the **coherence** of texts.
- Assumption: For every part of a coherent text, there is some function, some plausible reason for its presence, evident to readers.
- RST is intended to describe texts, rather than the processes of creating or reading and understanding them.
- Two spans of text (virtually always adjacent, but exceptions can be found) are related such that one of them has a specific role relative to the other.
- The claim span is called a **nucleus**, and the evidence span is called a **satellite**.

# RST Relations: Examples

---

<i>Relation</i>	<i>Nucleus</i>	<i>Satellite</i>
<b>background</b>	text whose understanding is being facilitated	text for facilitating understanding
<b>elaboration</b>	basic information	additional information
<b>contrast</b>	one alternate	the other alternate
<b>interpretation</b>	a situation	an interpretation of the situation
<b>justify</b>	text	information supporting the writer's right to express the text

# RST Diagram: Example

---

<b>1-5</b> <b>Preparation</b>				
1) Lactose and Lactase	<b>2-5</b> <b>Background</b>			
	<b>2-3</b> <b>Elaboration</b>		<b>4-5</b> <b>Contrast</b>	
	2) Lactose is milk sugar;	3) the enzyme lactase breaks it down.	4) For want of lactase most adults cannot digest milk.	5) In populations that drink milk the adults have more lactase, perhaps through natural selection.

# RST Corpus

---

- Rhetorical Structure Theory Discourse Treebank
- Selection of 385 Wall Street Journal articles from the Penn Treebank, annotated with discourse structure in the framework of RST
- Includes a number of humanly-generated extracts and abstracts associated with the original documents
- Distributed by the Linguistic Data Consortium (LDC)
- Purposes: discourse analysis, text generation, summarisation, machine translation, etc.

# RST: References

---

- William C. Mann and Sandra A. Thompson (1988): “Rhetorical Structure Theory: Toward a functional theory of text organization”. *Text*, 8(3):243-281.
- Maite Taboada and William C. Mann (2006, to appear): “Rhetorical Structure Theory: Looking back and moving ahead”. *Discourse Studies* 8(3).
- Maite Taboada and William C. Mann (2006, to appear): “Applications of Rhetorical Structure Theory”. *Discourse Studies*.

RST online: <http://www.sfu.ca/rst/>

# **Discourse Connectives in the Penn Discourse TreeBank**

---

# Penn Discourse TreeBank (PDTB)

---

- Annotation of **discourse connectives and their arguments**
- Characterisation of the semantic roles associated with the arguments of each type of connective; useful for syntax-semantics inferences in e.g. machine translation
- Built on top of the **Penn Treebank** and **PropBank**
- Connectives:  
subordinating, coordinating, adverbial, and implicit
- Procedure: **one connective at a time**;  
for implicit connectives the annotators read the entire text
- Evaluation: inter-annotator agreement (exact match)

# Discourse Connectives (1)

---

## Subordinating conjunctions:

- Clauses that syntactically depend on the main clause:
  - » **temporal** (such as *when, as soon as*)
  - » **concessive** (such as *because*)
  - » **purpose** (such as *so that, in order that*)
  - » **conditional** (such as *if, unless*)

*Because* [the drought reduced U.S. stockpiles], [they have more than enough storage space for their new crop], and that permits them to wait for prices to rise.

# Discourse Connectives (2)

---

## Coordinating conjunctions:

- Coordinating conjunctions are ones such as *and*, *but*, and *or*.
- Coordination of nominal, other non-clausal constituents, and VP-coordination are excluded.

[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].

# Discourse Connectives (3)

---

## Adverbial connectives:

- Sentence-modifying adverbs which express a discourse relation between two events or states, such as *however, therefore, then*
- Prepositional phrases which express similar binary relations, such as *as a result, in addition, in fact*

... [many analysts expected energy prices to rise at the consumer level too]. *As a result*, [many economists were expecting the consumer price index to increase significantly more than it did].

# Discourse Connectives (4)

---

## Implicit connectives:

- Relations between adjacent sentences that are not related by an explicit connective.
- Annotators are asked to provide an explicit connective that best describes the inferred relation.

... [The \$6 billion that some 40 companies are looking to raise in the year ending March 31 compares with only \$2.7 billion raised on the capital market in the previous fiscal year]. *IMPLICIT-(In contrast)* [In fiscal 1984 before Mr. Gandhi came to power, only \$810 million was raised].

# Arguments of Discourse Relations

---

- Annotation of **legal arguments** and **argument spans**
- Discourse relations hold between abstract objects.
- An argument contains at least one predicate along with its arguments:
  - » **single clause**
  - » **single sentence**
  - » **sequence of clauses and/or sentences**
  - » **combinations of both**
- Exceptions: **nominal phrases** that express an event or a state; **discourse deictics** that denote an event or a state

# Evaluation

---

- **Exact match criterion:** inter-annotator agreement in terms of agreement/disagreement on span identity for each token as a percentage of the pairs of spans that actually matched versus those that should have
- Binary values for any token within the arguments:
  - » 1 - identical textual selections
  - » 0 - non-identical selections
- Results: 90.2% agreement for the explicit connectives, 85.1% agreement for the implicit connectives
- 72% agreement on the type of explicit connective group for the implicit connectives

# Disagreement Analysis

---

- Error types:
  - » **partial overlap** (79%): some common span of text
  - » **no overlap** (5.6%): no overlap in the annotations
  - » **missing annotation** (13.5%): technical tool errors
  - » **unresolved** (1.9%): not in annotation guidelines
- Example of “higher verb” partial overlap:

[he knew the RDF was neither rapid nor deployable nor a force] - *even though* [it cost \$8 billion a year].

he knew [the RDF was neither rapid nor deployable nor a force] - *even though* [it cost \$8 billion a year].

# PDTB: References

---

- Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber (2004): “The Penn Discourse TreeBank”. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal.
- Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber (2004): “Annotating discourse connectives and their arguments”. In *Proceedings of the HLT/NAACL Workshop on “Frontiers in Corpus Annotation”*. Boston, MA.
- Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber (2005): “Experiments on sense annotations and sense disambiguation of discourse connectives”. In *Proceedings of the 4th Workshop on “Treebanks and Linguistic Theories”*. Barcelona, Spain.

PDTB online: <http://www.seas.upenn.edu/~pdtb/>

# **Anaphora and Coreference Annotation**

---

# Anaphora and Coreference Annotation

---

- **Coreference** is the reference in one expression to the same referent in another expression.
- **Anaphora** is coreference of one expression with its antecedent.

*A well-dressed man was speaking; he had a foreign accent.*

Examples for coreference annotation (among others):

- Prague Dependency Treebank
  - Coreferentially annotated corpus, Univ. of Wolverhampton
  - TüBa-D/Z, University of Tübingen
  - MATE
- Optimisation of machine learning NLP approaches

# Anaphora: References

---

- Ruslan Mitkov, Richard Evans, Constantin Orasan, Catalina Barbu, Lisa Jones, and Violeta Sotirova (2000): “Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies”. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference*, pp. 49-58. Lancaster, UK.
- Eva Hajičová, Jarmila Panevová, and Petr Sgall (2000): "Coreference in annotating a large corpus". In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pp. 497-500. Athens, Greece.
- Erhard Hinrichs, Sandra Kübler, Karin Naumann, Heike Telljohann, Julia Trushkina, and Heike Zinsmeister (2005): “Recent developments in linguistic annotations of the TüBa-D/Z Treebank”. Poster at the 27th Annual Meeting of the German Linguistic Society (Deutsche Gesellschaft für Sprachwissenschaft). Köln, Germany.

# **Prosodic Structure in the Kiel Corpus**

---

# Kiel Corpus of Spontaneous Speech

---

- Institute of phonetics and digital speech processing, University of Kiel, Germany
- Growing collection of read and spontaneous German
- Scenarios and data quantity:
  - » *appointment-scheduling*: 16 dialogues from 6 female, 8 male and 2 mixed pairs
  - » *video task / daily soap*: 6 dialogues from 4 female and 2 male pairs
- Duration: 80 minutes with ca. 13,000 consecutive words
- Application (example): natural text-to-speech

# Kiel Corpus: Data Collection

---

- **Appointment-Scheduling scenario:**

Two dialogue partners make various appointments on the basis of calendar sheets and academic time tables. When a speaker keeps a button pressed, his/her own speech signal is recorded, at the same time blocking the other speaker's channel.

- **Video Task / Daily Soap scenario:**

Similar but not identical video material is presented to two subjects sitting in separate rooms. After the presentations, the subjects discuss differences and similarities of what they have seen and heard. Both dialogue partners are recorded in parallel, natural interaction is not constrained.

# Kiel Corpus of Read Speech

---

- Institute of phonetics and digital speech processing, University of Kiel, Germany
- Growing collection of read and spontaneous German
- Contexts: train timetable queries, phonetically balanced material, two short stories
- Two speakers, male and female
- Current size: 624 isolated sentences with 4,932 word tokens and 1,673 word types
- Automatic extension with syntactic features, word frequency, and syllable boundaries (Brinckmann, 2005)

# Kiel Corpus: Annotation

---

- Prosodic annotation:
  - » lexical stress
  - » accent: location, type, degree, upstep
  - » intonation contour
  - » prosodic phrase boundaries
  - » pauses
- Phonetic annotation, such as plosive release phase, glottalisation, and nasalisation
- Deviations of the realised form from lexical phonemes, such as deletions, replacements, and insertions
- Orthography and punctuation marks
- Sentence and word boundaries

# Kiel Corpus: Example (1)

---

*Die Sonne lacht.* - `The sun is smiling.”

- oend: d i:+ z 'O n @ l 'a x t .
- kend: c: %d -h i:+ z 'O n @ l 'a x t -h .
- hend: 10361 #c:  
10361 ##%d  
10826 \$-h  
11070 \$i:+  
11830 ##z  
13019 \$'O  
14175 \$n  
14746 \$@ ...

# Kiel Corpus: Example (2)

---

*Am blauen Himmel ziehen die Wolken.*

- oend: Q a m+ b l 'aU @ n h 'l m @ l t s 'i: @ n d i:+  
v 'O l k @ n .
- kend: c: %Q -q a m+ b l 'aU @ n h 'l m @ %l t s 'i:  
@- n d -h i:+ v 'O l k -h @ n .
- hend: 6461 #c:  
6461 ##%Q  
7323 \$-q  
7323 \$a  
8534 \$m+  
9390 ##b ...

# Kiel Corpus: References

---

- Benno Peters (2005): “The database *The Kiel Corpus of Spontaneous Speech*”. In Klaus J. Kohler, Felicitas Kleber, Benno Peters (eds.): Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel *Prosodic Structures in German Spontaneous Speech*, no. 35a, pp. 1-6.
- Jonathan Harrington: “The phonetic analysis of speech corpora”. Manuscript. Parts available from <http://www.ipds.uni-kiel.de/PASCbook.html>
- Caren Brinckmann (2005): "The Kiel Corpus of Read Speech as a resource for prosody prediction in speech synthesis". In *Proceedings of the 2nd Baltic Conference on Human Language Technologies*. Tallinn, Estonia.

Kiel Corpus online:

<http://www.ipds.uni-kiel.de/forschung/kielcorpus.en.html>

# **Standardisation Projects: MATE and NITE**

---

# Corpus Annotation Resources

---

- Develop resources from scratch
- Acquire resources from previous projects and adapt to novel purposes
- Use results from various projects on annotation (and tools) and develop a standard for annotation, e.g. MATE and NITE, international projects with several partners

# The MATE Project

---

- MATE: multilevel annotation tools engineering
- Goal: facilitate the re-use of language resources by addressing the problems of **creating, acquiring, and maintaining language corpora**
- Tasks:
  - » development of a **standard for annotating resources**
  - » **provision of tools** which will make the processes of knowledge acquisition and extraction more efficient
- Focus: spoken language dialogue systems
- Multiple levels of spoken dialogue corpora: prosody, (morpho-) syntax, co-reference, dialogue acts, communicative difficulties, inter-level interaction

# MATE Workbench

---

- Annotation and exploration of relationships among different structures within a dialogue corpus
- Single interface to all of the basic functionalities which corpus annotators need
- Flexibility that different projects can provide different kinds of annotation and that information can be ported to applications
- Why this is hard: Corpus annotations are not necessarily hierarchically arranged. → XML for flexible representation of overlapping tag hierarchies.

# MATE: References

---

- Jean Carletta and Amy Isard (1999): "The MATE annotation workbench: User requirements". In *Proceedings of the ACL Workshop on "Towards Standards and Tools for Discourse Tagging"*.
- David McKelvie, Amy Isard, Andreas Mengel, Morten Baun Møller, Michael Grosse, and Marion Klein (2001): "The MATE workbench - An annotation tool for XML coded speech corpora". In *Speech Communication*, 33(1-2):97-112. Special issue on "Speech Annotation and Corpus Tools".

MATE online: <http://mate.nis.sdu.dk/>

# The NITE Project

---

- NITE: natural interactivity tools engineering
- Toolsets for multi-level, cross-level and cross-modality annotation, retrieval and exploitation of multi-party natural interactive human-human and human-machine dialogue
- Natural interactivity: enabling systems to exchange information with humans in the same ways in which humans exchange information with one another

# NITE Claims and Perspectives

---

- Claims:
  - » Ease and intuitiveness of human-human-system interaction could be improved tremendously through natural interaction.
  - » Pursuing natural interactivity leads to new families of applications which are not realised through traditional graphical interfaces with screen, mouse and keyboard.
- Perspectives:
  - » Toolset to analyse the complex coordination mechanisms used by humans to produce integrated natural interactive communication behaviour
  - » Corpora to constitute resources for programming or training machines to generate, recognise, and understand natural interactive communication

# NITE: References

---

- Niels Ole Bernsen, Laila Dybkjær, and Mykola Kolodnytsky (2002): “The NITE workbench - A tool for annotation of natural interactivity and multimodal data”. In Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, Spain.
- Niels Ole Bernsen, Laila Dybkjær, and Mykola Kolodnytsky (2002): “An interface for annotating natural interactivity”. In Jan van Kuppevelt and Ronnie W. Smith (eds.): “Current and New Directions in Discourse and Dialogue”. Kluwer, Dordrecht.

NITE online: <http://nite.nis.sdu.dk/>