MULI
Multilingual Information Structure

http://www.coli.uni-sb.de/cl/projects/muli/

TG 84-Projekt
Universität des Saarlandes

The MULI Team

- Fachrichtung 4.6 Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen
  - Stella Neumann, Christine Ortinau, Erich Steiner, Elke Teich
- Fachrichtung 4.7 Allgemeine Linguistik
  - Stefan Baumann, Thomas Blug, Caren Brinckmann, Irina Gogelgans, Silvia Hansen-Schirra, Geert-Jan M. Kruijff, Ivana Kruijff-Korbayová, Kerstin Kunz, Anja Moos, Hans Uszkoreit

Talk Outline

- Why corpus-based investigations of IS?
- What to investigate?
- Are suitable corpora available?
- Multi-level corpus annotation
  - Syntactic level
  - Intonation level
  - Discourse-entity level
- Integration for multi-level exploration and processing
- What next…

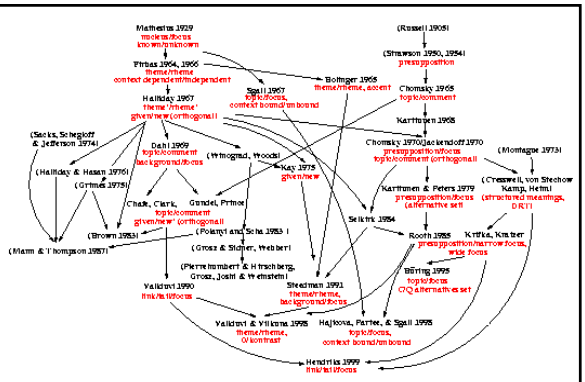
Information Structure

- (…) structural and semantic properties of utterances relating to
  - the discourse status of their content,
  - the actual and attributed attentional states of the discourse participants, and
  - the participants' prior and changing attitudes (knowledge, beliefs, intentions, expectations, etc.)

(Steedman/Kruijff-Korbayová 2003)

Information Structure

- Division of utterances or expressions into context-relating vs. context-affecting part(s)
  - reflecting "aboutness"
  - distinguishing among alternatives
- Diverse and underformalized terminology



Information Structure

- Division of utterances or expressions into context-relating vs. context-affecting part(s)
  - reflecting "aboutness"
  - distinguishing among alternatives
- Diverse and underformalized terminology
  - realization in different languages?
  - representation and interpretation?

What To Do?

- Corpus-based empirical study of information structure (IS)
- Explore correlations between various IS notions and aspects
  - where and how is content realized
- Specify and test theories about IS
  - role in discourse semantics (interpretation)
  - realization in different languages

Why Multilingual Perspective

- Typological perspective: how language properties lead to specific realization
- Use of different (combinations of) means to realize IS

Why Multilingual Perspective

- Sprachtypologische Eigenschaften führen zu sprachspezifischen Realisierungen
- Unterschiedliche Mittel können zur Informationsstrukturierung angewendet werden

Situation

- Lack of corpus data annotated with all suitable and necessary information
- Existing annotations scarce, disparate, and theory-specific
- Existing theories empirically inadequate: both too vague and too detailed
- Lack of annotation methodology

MULI Objectives

- Create corpus resources for IS research
- Investigate IS-relevant correlations
- Desiderata for annotation methodology
  - Theory-neutral notions
  - Robustness
  - Cross-linguistic applicability
  - Genre/Register independence
  - Multiple layers

MULI IS Annotation Approach

- Multiple levels of annotation
  - Syntax
  - Intonation
  - Discourse
- Not annotation of IS directly, but rather
- Annotation of features relevant for IS
  - at each level
  - in each language

What to investigate...

Syntax

- IS-relevant aspects of realization
  - Positioning
  - Ordering (words, phrases, clauses)
  - Marked syntactic constructions
  - Morphological marking (in some languages)
  - Definiteness marking (in some languages)
- Cross-linguistic variation in repertoire and use as means of IS realization...

Marked Syntactic Structures

A mechanic is fixing a car. – „congruent“

There´s a mechanic fixing a car. – thetic

It´s a mechanic that's fixing a car. – clefts

It´s a car that a mechanic is fixing.

A car is being fixed by a mechanic. – passive

A mechanic fixed the car yesterday. – order

Yesterday a mechanic fixed the car.

(Crystal 1995)

Differences Between Languages

Yesterday, a mechanic fixed the car.

A mechanic fixed the car yesterday.

The car, a mechanic fixed yesterday.

Gestern hat ein Mechaniker das Auto repariert.

Ein Mechaniker hat gestern das Auto repariert.

Ein Mechaniker hat das Auto gestern repariert.

Das Auto hat gestern ein Mechaniker repariert.

Das Auto hat ein Mechaniker gestern repariert.

Intonation

- IS-relevant aspects of realization
  - Position and type of boundary tones
  - Position and type of accents
  - Position and size of phrase boundaries
  - Pitch range, rythm, speed ...
- Cross-linguistic variation in repertoire and use as means of IS realization...

Discourse

- Discourse phenomena relevant to IS:
  - Reference and various discourse referent properties
    - givenness /familiarity/identifiability
    - delimitation (e.g., specificity), quantification
    - semantic sorts
  - Anaphoric relations
  - Rhetorical relations
- Information distribution and density
- Cross-linguistic variation in interplay of discourse phenomena with IS...

Where to investigate...

Existing Linguistic Resources

- Bad news
  - There is no corpus where all aspects relevant for IS are annotated
- Good news
  - There are corpora that have some of it
  - Emerging methodology, standards, tools

Existing Linguistic Resources

- Syntactic treebanks, sometimes with semantics
  - Surface syntax, isolated sentences
  - Argument structure, dependency relations, frames
  - Quantifier scopes
- Corpora with discourse annotation
  - Anaphoric relations
  - Rhetorical relations
  - Dialogue moves
- Speech corpora with intonation annotation

Interesting Existing Resources

- IS-relevant information
- Multiple levels
- Extendability

PennTreebank

- English newspaper texts
- Constituent-based syntax
- Dependency-based syntax (conversions)
- Semantics: PropBank
- Anaphoric relations
- Discourse relations:
  - Daniel Marcu´s RST corpus
  - D-LTAG project at Upenn

Prague Dependency Treebank

- Czech newspaper texts (also English from PTB)
- Dependency-based syntax
  - Analytical level (surface)
  - Tectogrammatical level (deep)
- Coreference relations
- Topic-focus articulation
  - contextual boundness
  - contrast

Verbmobil corpus

- English and German spoken dialogue data
- Dialogue act annotation
- Intonation
- LinGO-Redwoods treebank (Stanford)
  - English
  - HPSG syntax
- ROSIE project (Edinburgh)
  - SDRT-based discourse annotation

NeGra / TiGer Treebank

- German newspaper text
- Dependency-based surface syntax
- Coreference relations
- Frames (following the FrameNet project)
- TiGer Registry: conversion of various corpora into TiGer XML format

MULI Corpus

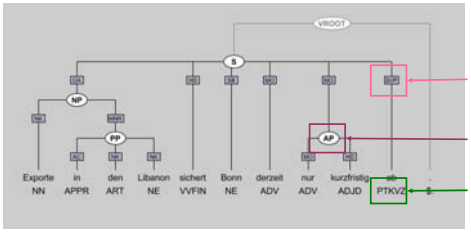
- German: extract from TiGer-Treebank
  - economy news from Frankfurter Rundschau
  - cca. 3.500 words in cca. 250 sentences
- English: extract from Penn Treebank
  - Wall Street Journal articles
  - cca. 7.000 words in cca. 320 sentences

Example Text

Rexrodt hält es daher für wichtig, den Libanon, aber auch Syrien in den Friedensprozeß einzubeziehen. Beeindruckt zeigte er sich vom voranschreitenden Aufbau in der teilweise stark zerstörten libanesischen Hauptstadt Beirut. Den in diesem Land vorhandenen Unternehmergeist lobte der Minister als eine ``milliarden-schwere Infrastruktur´´. ``Da steckt Musik drin´´, sagte er und ermunterte die deutsche Wirtschaft zu einem verstärkten Engagement. Dem stehen, wie libanesische Gesprächspartner beklagten, beschränkte Ausfuhrgarantien entgegen. Exporte in den Libanon sichert Bonn derzeit nur kurzfristig ab. Bei aussichtsreichen mittel- und lang-fristigen Vorhaben versprach Rexrodt nun eine Einzelfallprüfung. Skeptisch beurteilte er von palästinensischer Seite gewünschte Direktinvestitionen im Gaza-Streifen und Westjordanland.

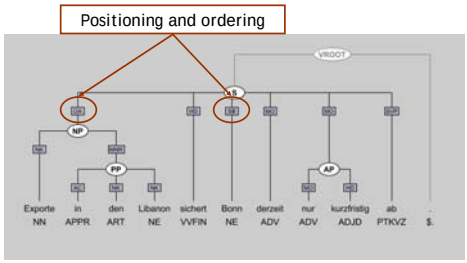
Syntax Level

Syntax in the TiGer-Treebank



+ Morphologie

Syntax in the TiGer-Treebank



Additional Syntax Annotation

- Presence of marked structures in clauses
  - Cleft
  - Pseudo-cleft
  - Extraposition
  - Fronting
  - Expletive 'es' / 'There'-insertion
  - Diathesis: passive
- Tool: XMLspy (XML editor)

Syntax annotation

- Clefting
  - Julie buys her vegetables at the market
  - It's Julie/her vegetables/at the market who/that/where…
- Pseudo-clefting (cf. Funktion des Passiv); + reversed pseudo-clefting
  - Julie buys her vegetables at the market.
  - What Julie buys at the market are her vegetables.
  - Her vegetables are what Julie buys at the market.
- Passive

Syntax annotation

- Fronting
  - Julie buys her vegetables at the market.
  - Her vegetables Julie buys at the market.
- Extraposition
  - What you say doesn't matter.
  - It doesn't matter what you say.
- There-insertion (existential there); D: es
  - Someone was knocking at the door.
  - There was someone knocking at the door.

Examples from the TiGer-Treebank

- Clefting/pseudo-clefting: keine
- Fronting:
  - Beeindruckt zeigte er sich vom voranschreitenden Aufbau in der teilweise stark zerstörten libanesischen Hauptstadt Beirut.
  - Exporte in den Libanon sichert Bonn derzeit nur kurzfristig ab.
  - Hintergrund sind die geschäftlichen Einschränkungen und Imageschäden…

Examples from the TiGer Treebank

- Extraposition:
  Es paßt nicht in die Mentalität dieser Leute, zu schmuggeln,
  Es ist ja nicht so, daß die Europäer plötzlich die DM regieren
- There-insertion: none (but: "es gibt")
  Er hoffe, daß es nicht auch „in der SPD Leute gibt, die meinen, ... ".
- Passive
  Rentner und Sparer müßten überzeugt werden,...

Example Text

Rexrodt hält es daher für wichtig, den Libanon, aber auch Syrien in den Friedensprozeß einzubeziehen. ***Beeindruckt*** *zeigte er sich vom voranschreitenden Aufbau in der teilweise stark zerstörten libanesischen Hauptstadt Beirut.* ***Den in diesem Land vorhandenen Unternehmergeist*** *lobte der Minister als eine ``milliarden-schwere Infrastruktur´´. ``Da steckt Musik drin´´, sagte er und ermunterte die deutsche Wirtschaft zu einem verstärkten Engagement.* ***Dem*** *stehen, wie libanesische Gesprächspartner beklagten, beschränkte Ausfuhrgarantien entgegen.* ***Exporte*** *in den Libanon sichert Bonn derzeit nur kurzfristig ab.* ***Bei aussichtsreichen mittel- und lang-fristigen Vorhaben*** *versprach Rexrodt nun eine Einzelfallprüfung.* ***Skeptisch*** *beurteilte er von palästinensischer Seite gewünschte Direktinvestitionen im Gaza-Streifen und Westjordanland.*

Contrastive Comparison

	DE		EN	
	absolute	in %	absolute	in %
Clauses	328		776	
Passive	44	13,41	88	11,34
n/a	297	90,55	746	96,13
cleft	0	0,00	1	0,13
pseudo-cleft	0	0,00	2	0,26
reversed-pseudo-cleft	0	0,00	0	0,00
fronting	29	8,84	8	1,03
Extraposition	2	0,61	5	0,64
"es" / "*there*"	0	0,00	14	1,80

Further steps

- Additional annotation (or extraction from treebank):
  - Other aspects of ordering in German, e.g, field topology
  - Focus adverbs, ect.
- Interpretation of findings in terms of information distribution w.r.t. other levels
  - Intonation in sentences with syntactically marked constructions
  - Information status and anaphoricity of fronted expressions

Intonation Level

Intonation Annotation

- Read texts from the corpus
- Annotation according to GToBI
  - Position and type of boundary tones
  - Position and type of accents
  - Position and size of phrase boundaries
- Tool: EMU (speech analysis tool)

Intonation Analysis Example

Dem stehen, wie libanesische Gesprächspartner beklagten, beschränkte Ausfuhrgarantien entgegen.

Exporte in den Libanon sichert Bonn
L+H* L*+H H-
derzeit nur kurzfristig ab.
H+ L* L-%

Bei aussichtsreicheren mittel- und langfristigen Vorhaben versprach Rexrodt nun eine Einzelfallprüfung.

Intonation Analysis

- So far only very preliminary observations, e.g.,
  - Fronted constituents
    - intonation phrases
    - rising accents
- Problems
  - Multiple accents per word
  - Orthography vs. phonology

Discourse Level

Discourse Level Annotation

- Annotation of expressions with
  - Discourse referents
  - Referent properties
    - Extentional vs. intensional reference
    - Semantic sort
    - Unique, existential or variable delimitation
    - Countability and quantification
    - Information status (Prince´s familiarity taxonomy)
  - Anaphoric relations
    - Identity of reference or sense; various types of bridging
- Tool: MMAX (text annotation tool)

Example Text

Rexrodt hält es daher für wichtig, den Libanon, aber auch Syrien in den Friedensprozeß einzubeziehen. Beeindruckt zeigte er sich vom voranschreitenden Aufbau in der teilweise stark zerstörten libanesischen Hauptstadt Beirut. Den in diesem Land vorhandenen Unternehmergeist lobte der Minister als eine ``milliarden-schwere Infrastruktur´´. ``Da steckt Musik drin´´, sagte er und ermunterte die deutsche Wirtschaft zu einem verstärkten Engagement. Dem stehen, wie libanesische Gesprächspartner beklagten, beschränkte Ausfuhrgarantien entgegen. Exporte in den Libanon sichert Bonn derzeit nur kurzfristig ab. Bei aussichtsreichen mittel- und langfristigen Vorhaben versprach Rexrodt nun eine Einzelfallprüfung. Skeptisch beurteilte er von palästinensischer Seite gewünschte Direktinvestitionen im Gaza-Streifen und Westjordanland.

Example Text

Rexrodt hält es daher für wichtig, den Libanon, aber auch Syrien in den Friedensprozeß einzubeziehen. Beeindruckt zeigte er sich vom voranschreitenden Aufbau in der teilweise stark zerstörten libanesischen Hauptstadt Beirut. Den in diesem Land vorhandenen Unternehmergeist lobte der Minister als eine ``milliarden-schwere Infrastruktur´´. ``Da steckt Musik drin´´, sagte er und ermunterte die deutsche Wirtschaft zu einem verstärkten Engagement. Dem stehen, wie libanesische Gesprächspartner beklagten, beschränkte Ausfuhrgarantien entgegen. Exporte in den Libanon sichert Bonn derzeit nur kurzfristig ab. Bei aussichtsreichen mittel- und langfristigen Vorhaben versprach Rexrodt nun eine Einzelfallprüfung. Skeptisch beurteilte er von palästinensischer Seite gewünschte Direktinvestitionen im Gaza-Streifen und Westjordanland.

How Much Work Is It?

- It´s hard...
  - Also designing annotation methodology and schemes
- And quite time-consuming
  - Intonation: 30 min/sent. (incl. discussion)
  - Discourse: 10 min/sent. (incl. thinking about scheme)
    - Create markables and links, assign properties
  - Syntax: 5 min/sent. --only addt´l annotation
    - Segmenation into clauses, classification; direct XML editing!
- But it must be worth it!

Interactions Between Levels: Preliminary Observations

- Dem stehen, wie libanesische Gesprächspartner beklagten, beschränkte Ausfuhrgarantien entgegen.
- Exporte in den Libanon sichert Bonn derzeit nur kurzfristig ab.
- Bei aussichtsreicheren mittel- und langfristigen Vorhaben versprach Rexrodt nun eine Einzelfallprüfung.

Example analysis

Dem stehen, wie libanesische Gesprächspartner beklagten, beschränkte Ausfuhrgarantien entgegen.

Exporte in den Libanon sichert Bonn

derzeit nur kurzfristig ab.

Bei aussichtsreicheren mittel- und langfristigen Vorhaben versprach Rexrodt nun eine Einzelfallprüfung.

Example analysis

- Dem stehen, wie libanesische Gesprächspartner beklagten, beschränkte Ausfuhrgarantien entgegen.
- Exporte in den Libanon sichert Bonn
- derzeit nur kurzfristig ab.
- Bei aussichtsreicheren mittel- und langfristigen Vorhaben versprach Rexrodt nun eine

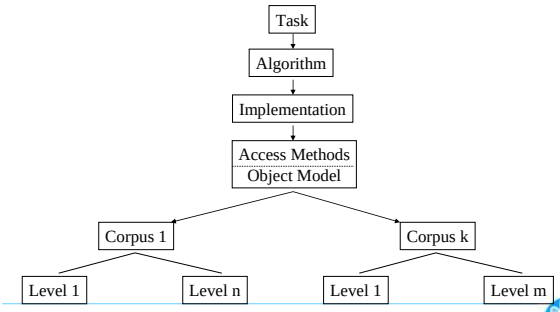
Integration

- Annotation on multiple levels, specific tools
- Integration of annotation across levels needed
  - for exploration (viewing)
  - for processing, e.g., discovery of correlations, tagging
- How to integrate?
  - convertion of "native" data formats into a common data format
  - interface(s) for accessing "native" format(s)
- Important issues of multi-level annotation integration
  - Stand-off annotation
  - Alignment of base data

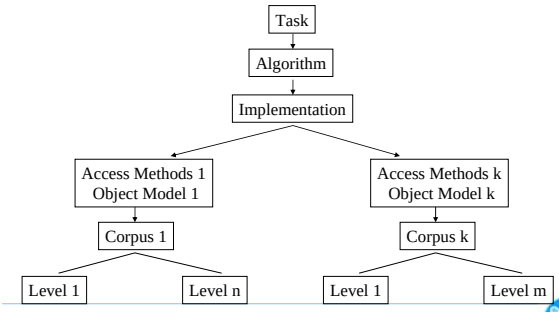
MMAX Discourse API (Java)

- Concepts
  - Markables (and attributes)
  - Links
- Interface to corpora annotated in the MMAX data format (at multiple levels)
  - Methods for accessing markables and links at different levels of annotation
  - Parses MMAX files, builds data structures representing the corpus and provides a *Discourse* object

Integration: With Conversion



Integration: With Interface



Integration: Discourse Level

- Concepts:
  - Markables:
    - expressions introducing discourse referents
    - markable attributes: properties of referents
  - Links:
    - anaphoric relationships
- Annotation in MMAX data format
- MMAX Discourse object

Integration: Syntax

- Concepts:
  - Markables:
    - expressions introducing syntactic-tree nodes
    - markable attributes: synt. features and edge labels
  - Links:
    - edges in tree
- Annotation in Tiger XML data format
- Conversion to MMAX (syntactic markable level)
- Tiger API (creates Corpus object)

Tiger API (Java)

- Interface to the TIGER corpus and other corpora encoded in TIGER-XML
  - Access the structure of a corpus given as a TIGER-XML file
  - Methods for traversing the syntax trees, accessing single elements such as sentences, nodes, non-terminals, terminals and their attributes are provided.
  - Parses a given TIGER-XML file, builds data structures representing the corpus and provides a *Corpus* object

Integration: Intonation

- Concepts:
  - Markables:
    - words and intonation phrases
    - markable attributes: accent types, boundary tones and break sizes
  - Links: none
- Annotation in EMU data format
- No conversion or interface yet

What Next

- Annotation completion, polishing up, reliability evaluation
- Preliminary observations (mainly manual?)
- Integration
- Further investigation (computational)
- Annotation bootstrapping solutions

⇒ Follow-up project

Thank You!

http://www.coli.uni-sb.de/cl/projects/muli/