Bergen, December 5-6 2007
Workshop
"Unified Linguistic Annotation – Transcontinental Perspectives"

Broad-coverage LFG Parsing:
Beyond Parallel Grammar Development

Jonas Kuhn, University of Potsdam, Germany

Question for this presentation

- What role do/should treebanks play in advanced work with deep linguistic grammars/parsers?

(Motivation/Hope:)

- Hand annotation is expensive
  - Annotated resources should be shared as much as possible
- Facilitate comparison across frameworks/languages

Question for this presentation

- What role do/should treebanks play in advanced work with deep linguistic grammars/parsers?
- Examples of work with deep linguistic grammars
  - LFG ParGram project: grammar development
  - "Beyond grammar development"
    - ParGram grammars in Syntax-based Statistical MT
    - ParGram grammars in linguistic research on unannotated corpora
    - Semantic construction on top of ParGram grammars

Question for this presentation

- What role do/should treebanks play in advanced work with deep linguistic grammars/parsers?
- "Treebank": Structurally annotated linguistic data, involving human assessment
  - Broad view, leaving open:
    - Level of annotation
    - Mode of data selection
    - Exhaustiveness/fixedness of guidelines
    - Scale of human assessment
    - Role played by "treebank"

Degrees of freedom…

- Level of annotation
  - morphological, syntactic (dependency/grammatical relations, phrase structure), argument structure/lexical semantics, deep(er) semantics
- Mode of data selection
  - corpus data (representativeness?), constructed data (= testsuite), filtered corpus data
- Exhaustiveness/fixedness of guidelines
  - forced decision, postponing decisions, “dynamic” scheme
- Scale of human assessment
  - complete manual annotation, semi-automatic annotation, selection among parser results
- Role played by “treebank”
  - objective (task-independent?) evaluation
  - development data of various kinds

Question for this presentation

- What role do/should treebanks play in advanced work with deep linguistic grammars/parsers?

Skip ahead to Conclusions…

(Potentially) controversial points

- “Unified annotation” may be hard to achieve in many areas that involve use of annotated linguistic data
- Few NLP “modules” can be reasonably assessed in a task-independent way
- Many uses of structurally annotated corpus data require a scale that is unrealistic to achieve by manual annotation

Consequences?

- Strive for comparative annotation in gold standards for evaluation
  - Overlap in annotated data
- Acknowledge the need for various different (and often flexible) annotation schemes in NLP research
- Facilitate re-use of those resources
  - Establish high standards for annotation guidelines, documentation of annotation efforts, meta-evaluation
  - Careful book-keeping over available annotations
  - Smooth integration of formats
- Important goal: Transparent automatic annotation tools with confidence assessment

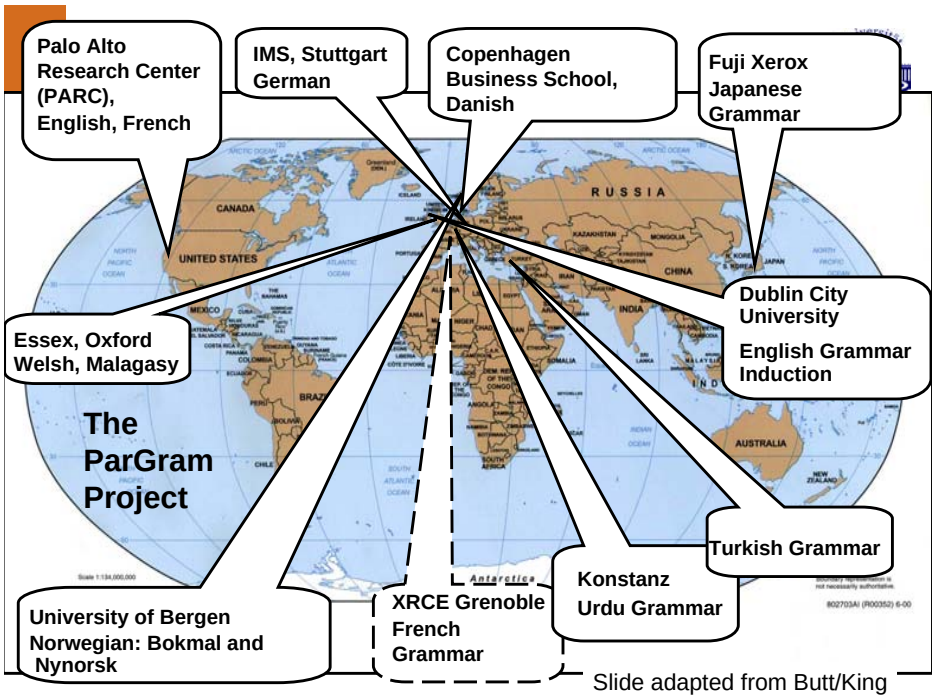
Question for this presentation

- What role do/should treebanks play in advanced work with deep linguistic grammars/parsers?

Back to the main part…

Use of "treebanks"

- Examples of work with deep linguistic grammars
  - LFG ParGram project: grammar development
  - "Beyond grammar development"
    - ParGram grammars in Syntax-based Statistical MT
    - ParGram grammars in linguistic research on unannotated corpora
    - Semantic construction on top of ParGram grammars



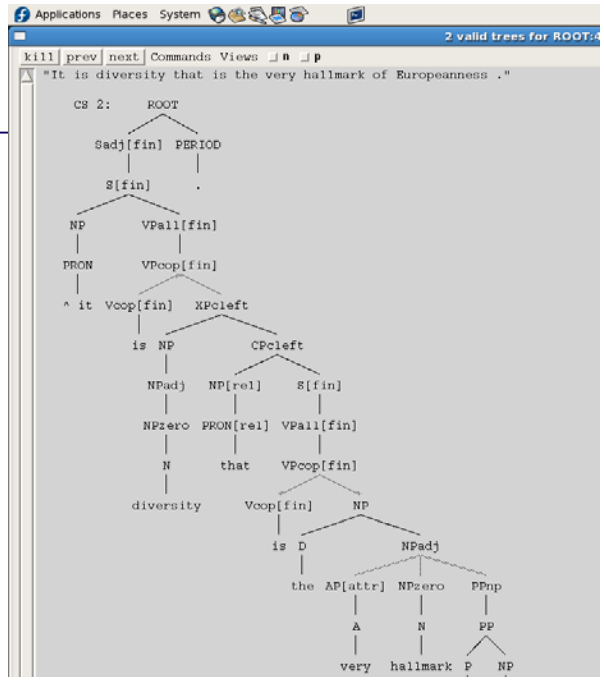
Slide adapted from Butt/King

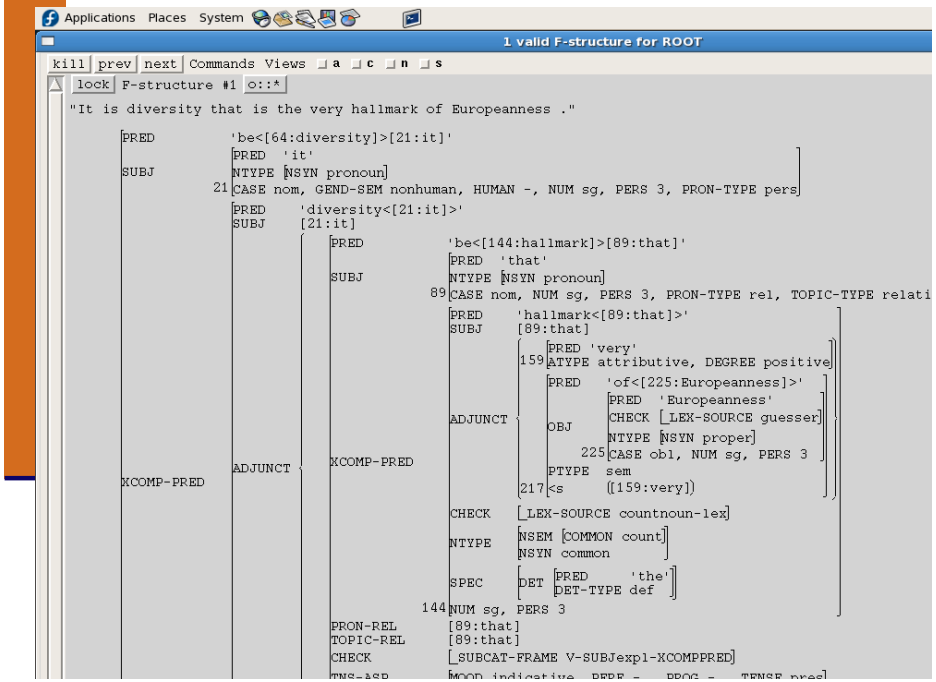
ParGram

- LFG
- XLE

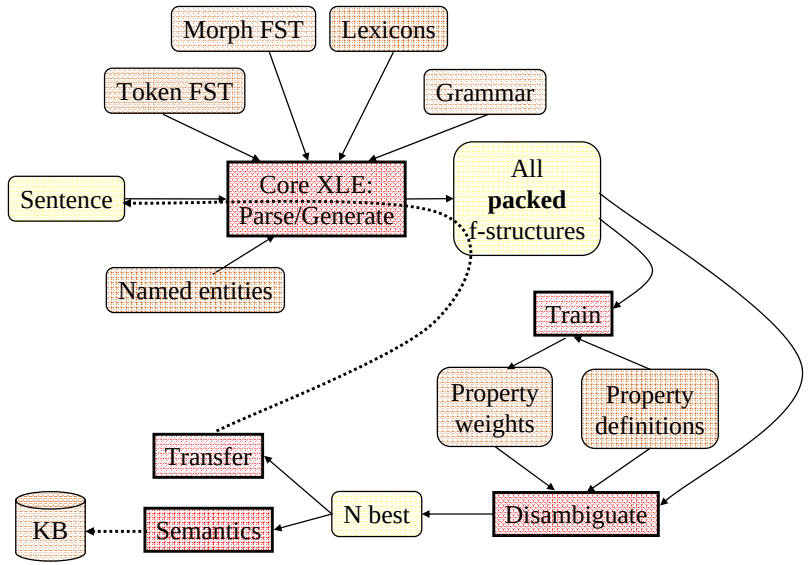
Development environment,

Parser/ generator





XLE related language components



Slide adapted from Butt/King

Grammar development

- ParGram Methodology (personal view)
  - **Grammar development**
    - Parallel
    - Theory-based
    - Data-driven

- Syntax
- Morphology
- Lexicon

B r o a d C o v e r a g e

English | French | German | Norwegian | Japanese | …

Lexical-Functional Grammar

Data in grammar development

- Driving the agenda
  - Coverage of grammatical phenomena
  - Coverage of corpus data (in envisaged application domain)
  - Parallelism across grammar
- Ambiguity management, profiling
  - Advanced grammar development
  - Machine learning for disambiguation task
- Comparative quality assessment

Test suites

Plain corpora

Comparative annotated test suites

Annotated corpora/test suites

Approximative use of existing **large** treebanks + conversion, Semi-automatic annotation

Use of existing treebanks + conversion

Data in grammar development

- Driving the agenda
  - Coverage of grammatical phenomena
  - Coverage of corpus data (in envisaged application domain)
  - Parallelism across grammar
- Ambiguity management, profiling
  - Advanced grammar development
  - Machine learning for disambiguation task
- Comparative quality assessment

Flexibility in "annotation guidelines" crucial during development

Purely corpus-oriented representativeness less central; filtering legitimate

Representativeness very central; task independence (!?)

Data in grammar development

- In relatively early stages of developing a particular grammar, reusability of internal annotated data may be limited
- For a mature grammar, "multi-purpose treebanking" is realistic
  - LFG representations are compatible with both phrase structure and dependency oriented annotation
  - Keeping track of original purpose/circumstances behind annotation is however important (e.g., parser-based semi-automatic annotation)
  - Parallelism still justifies use of constructed/modified data (besides random sampled corpus data)

Use of "treebanks"

- Examples of work with deep linguistic grammars
  - LFG ParGram project: grammar development
  - **"Beyond grammar development"**
    - ParGram grammars in Syntax-based Statistical MT
    - ParGram grammars in linguistic research on unannotated corpora
    - Semantic construction on top of ParGram grammars

Use of "treebanks"

- Examples of work with deep linguistic grammars
  - LFG ParGram project: grammar development
  - "Beyond grammar development"
    - **ParGram grammars in Syntax-based Statistical MT**
    - ParGram grammars in linguistic research on unannotated corpora
    - Semantic construction on top of ParGram grammars

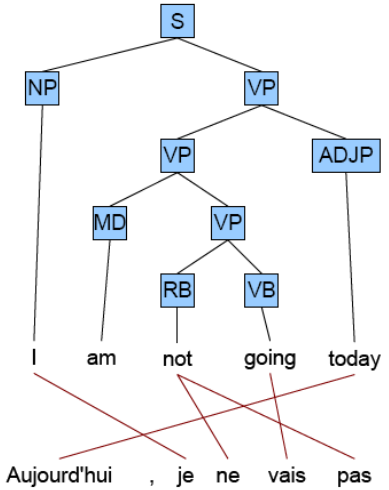
Syntax-based StatMT

- PTOLEMAIOS project (Saarbrücken/Potsdam)
  - Training on parallel corpus
    - Source language parse
      - Phrase structure
        + Statistical word alignment
        → Extract transduction rules
        [following Galley/Hopkins/Knight/Marcu 2004]
      - Feature structures
        → Train cascade of classifiers for controlling application of rules
    - Target language parse (in training)
      - Factorization by morphological features

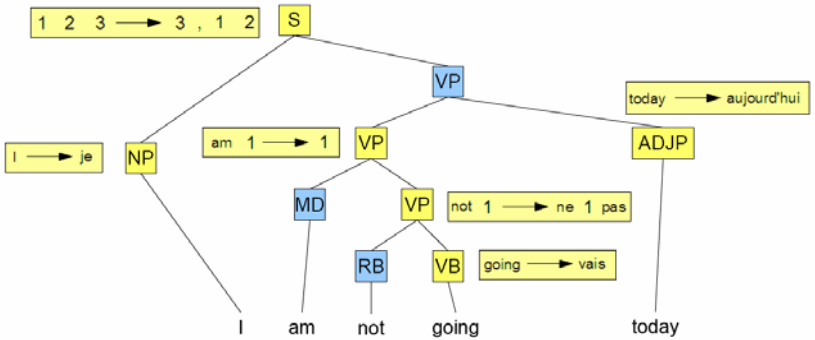
Source language analysis
(ParGram-based: trees + feature structures)

(Statistical) word alignment

Target language: morphological analysis



- Transduction rules



- Rich syntactic representations provided by LFG grammars are a good basis for training the classifier cascade
  - Inducing a generation grammar for target language, relative to source language

"Treebanks" in StatMT?

- No use at all? [manual annotation]
  - Automatic parsing is crucial (especially in application of the translator, but also for generating training data)
- Evaluation of parser quality can be very useful for error analysis, improved feature design etc.
  - Differentiated multi-purpose view on treebank may cater for this
  - Parallel treebank with correspondence annotation would be particularly helpful

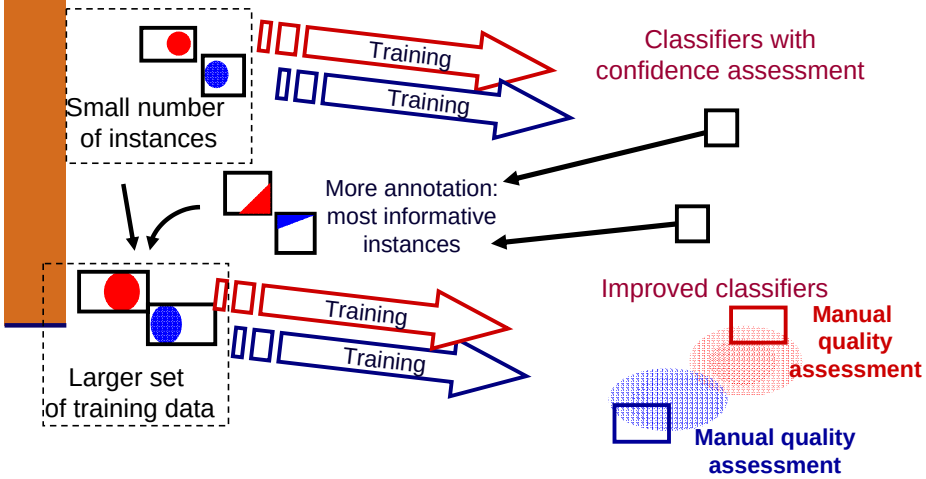
Use of “treebanks”

- Examples of work with deep linguistic grammars
  - LFG ParGram project: grammar development
  - “Beyond grammar development”
    - ParGram grammars in Syntax-based Statistical MT
    - **ParGram grammars in linguistic research on unannotated corpora**
    - Semantic construction on top of ParGram grammars

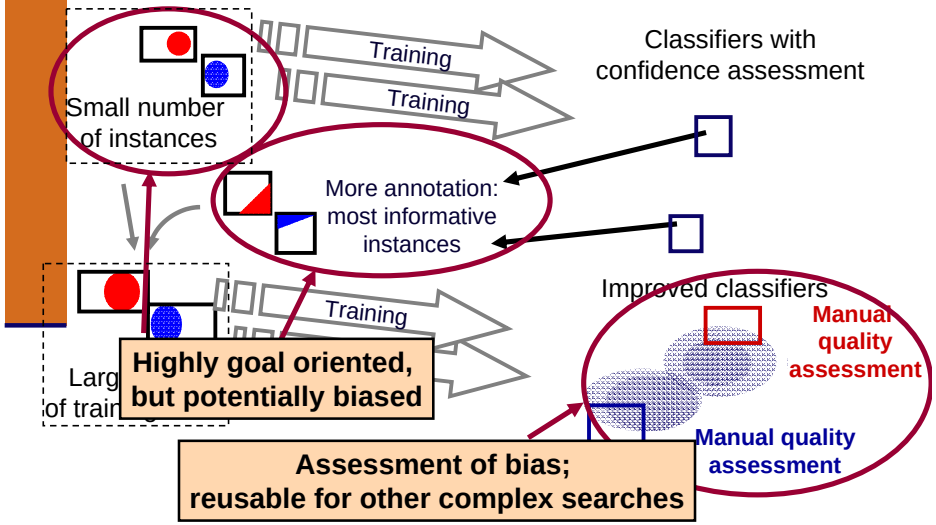
Research on unannotated corpora

- Project D4 within
  SFB 632 on Information Structure (Potsdam/ Humboldt University, Berlin)
  - Many relevant phenomena are too rare to find enough instances in hand-annotated treebanks
  - Exploit
    - NLP tools (and existing annotations) and
    - expertise of linguistic researchers (and their overlap of interests) in an interactive approach to corpus exploration/partial annotation
  - Apply ideas from Active Learning
  - Use parallel corpora to bridge across languages

Interactive corpus exploration

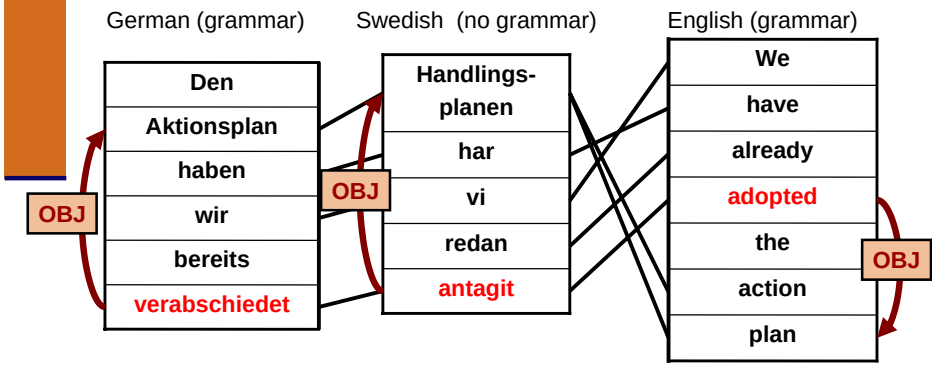


Manual effort



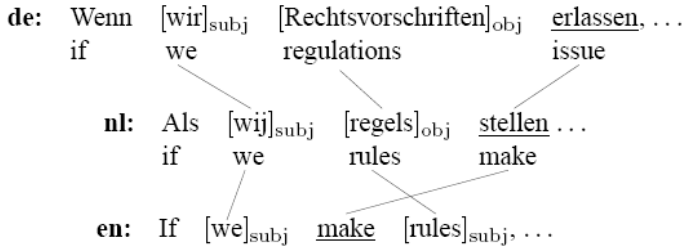
Pilot study

- "Multi-parallel annotation projection"
  - Train argument-head classifier for target language, exploiting two parallel grammars



Pilot study

- Current experiments for Dutch as a target language
  - Europarl-based, easy generalization to all other languages



Pilot study

- Maximum Entropy classifiers for individual argument-head candidate pairs
  (86.2% precision and 60.8% recall in first small pilot study – without Active Learning)
- "Global" classifier, deciding on all arguments of a given head (verb) under development
- Next steps:
  - (Inter-)active training
  - Combination of classifiers for various linguistic properties/dimensions

Role of "treebanks"?

- Hand annotated parallel treebank would be of high value in various contexts
  - Core data for initial corpus research
  - Seed data for bootstrapping
  - Quality assessment (bias checking)
- Exploit existing annotated resources wherever possible

Role of "treebanks"?

- Combination of various types of annotated data seems reasonable
  - (Representative!?) corpus data
  - Elicited production data (available from a broad typological spectrum for information structural phenomena)
  - Real-life translations
  - Literal translations
  - Parallel constructed "test suites" of corresponding constructions (which need not be translational equivalents)
- Important: differentiated multi-purpose view on "treebanks"

Use of "treebanks"

- Examples of work with deep linguistic grammars
  - LFG ParGram project: grammar development
  - "Beyond grammar development"
    - ParGram grammars in Syntax-based Statistical MT
    - ParGram grammars in linguistic research on unannotated corpora
    - **Semantic construction on top of ParGram grammars**

Semantic construction

- Various frameworks under discussion
  - Glue language semantics (following work by Dalrymple and colleagues)
  - LFG-based MRS construction (LOGON project)
  - "Transfer semantics", exploiting XLE's term rewrite transfer system
    - Data-driven engineering approach
    - Broad-coverage system for English developed at PARC, using lexical resources such as WordNet
    - Thanks to the ParGram idea, adoption to other languages is straightforward (adoption to Japanese by Hiroshi Masuichi)

Semantic Representation

Someone failed to pay

```
in_context(t, past(fail22))
in_context(t, role(Agent, fail22, person1))
in_context(t, role(Predicate, fail22, ctx(pay19)))
in_context(ctx(pay19), cardinality(person1, some))
in_context(ctx(pay19), role(Agent, pay19, person1))
in_context(ctx(pay19), role(Recipient, pay19, implicit_arg94))
in_context(ctx(pay19), role(Theme, pay19, implicit_arg95))

lex_class(fail22, [vnclass(unknown), wnclass(change),
                   temp-rel, temp_simul, impl_pn_np, prop-attitude])
lex_class(pay19, [vnclass(unknown), wnclass(possession)])),
word(fail22, fail, verb, 0, 22, t, [[2505082], [2504178], …, [2498138]])
word(implicit_arg:94, implicit, implicit, 0, 0, ctx(pay19), [[1740]])
word(implicit_arg:95, implicit, implicit, 0, 0, ctx(pay19), [[1740]])
word(pay19, pay, verb, 0, 19, ctx(pay19),
                   [[2230669], [1049936], …, [2707966]])
word(person1, person, quantpro, 0, 1, ctx(pay19),
                   [[7626, 4576, …, 1740]])
```

Slide adapted from Butt/King

Abstract Knowledge Repres.

Someone failed to pay

Conceptual Structure:

subconcept(fail22, [[2:2505082], [2:2504178], …, [2:2498138]])
role(Agent, fail22, person1)
subconcept(person1, [[1:7626, 1:4576, …, 1:1740]])
role(cardinality_restriction, person1, some)
role(Predicate, fail22, ctx(pay19))
subconcept(pay19, [[2:2230669], [2:1049936], …, [2:2707966]])
role(Agent, pay19, person1)

Contextual Structure:

context(t) context(ctx(pay19))
context_lifting_relation(antiveridical, t, ctx(pay19))
context_relation(t, ctx(pay19), Predicate(fail22))
instantiable(fail22, t)
uninstantiable(pay19, t)
instantiable(pay19, ctx(pay19))

Temporal Structure:

temporalRel(startsAfterEndingOf, Now, fail22)
temporalRel(startsAfterEndingOf, Now, pay19)

Slide adapted from Butt/King

Role of "treebanks"?

- Towards Parallel Semantics
- Similar flexibility requirements as within early syntactic grammar development seem to apply
  - Even in more "mature stages", granularity of semantic analysis will presumably remain controversial
- Task-independent status of annotation much less clear than with syntactic parsing (or frame-oriented annotation of lexical semantics)
- Differentiated picture of annotation very important
- Comparative cross-linguistic data sets will be highly interesting

Use of "treebanks"

- Examples of work with deep linguistic grammars
  - LFG ParGram project: grammar development
  - "Beyond grammar development"
    - ParGram grammars in Syntax-based Statistical MT
    - ParGram grammars in linguistic research on unannotated corpora
    - Semantic construction on top of ParGram grammars

Question for this presentation

- What role do/should treebanks play in advanced work with deep linguistic grammars/parsers?

Now again the (tentative) Conclusions…

Need annotations of various kinds

- Level of annotation
  - morphological, syntactic (dependency/grammatical relations, phrase structure), argument structure/lexical semantics, deep(er) semantics
- Mode of data selection
  - corpus data (representativeness?), constructed data (= testsuite), filtered corpus data
- Exhaustiveness/fixedness of guidelines
  - forced decision, postponing decisions, "dynamic" scheme
- Scale of human assessment
  - complete manual annotation, semi-automatic annotation, selection among parser results
- Role played by "treebank"
  - objective (task-independent?) evaluation
  - development data of various kinds

(Potentially) controversial points

- "Unified annotation" may be hard to achieve in many areas that involve use of annotated linguistic data
  - Development data both in symbolic and machine-learning work will often reflect peculiarities of the approach taken
- Few NLP "modules" can be reasonably assessed in a task-independent way
  - Syntactic parsing may be atypical in this respect
- Many uses of structurally annotated corpus data require a scale that is unrealistic to achieve by manual annotation

Consequences?

- Strive for comparative annotation in gold standards for evaluation
  - Overlap in annotated data
  - Special case: parallel corpora – translational overlap would be extremely helpful
  - Where possible break up complex annotation in component aspects
    - More transparent for quality assessment
    - Recombinable

Consequences?

- Acknowledge the need for various different (and often flexible) annotation schemes in NLP research
- Facilitate re-use of those resources
  - Establish high standards for annotation guidelines, documentation of annotation efforts, meta-evaluation
  - Careful book-keeping over available annotations
  - Smooth integration of formats
- Important goal: Transparent automatic annotation tools with confidence assessment

Acknowledgements

- Work in the ParGram project
  - PARC, Bergen, IMS Stuttgart, Fuji Xerox …
- Ongoing work from Potsdam group
  - PTOLEMAIOS project (on StatMT):
    - Mark Hopkins
  - SFB 632 on Information Structure, Project D4:
    - Bettina Schrader
    - Kathrin Spreyer
    - Gerlof Bouma
  - Ongoing Master thesis (on Semantic Construction):
    - Sina Zarrieß