Networks of Linguistic Annotation

The Linguist's Web

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

Anette Frank
Computational Linguistics Department
University of Heidelberg

Overview

Grammar-based linguistic annotation

Treebanks: Varieties and Usages

Monolingual – Multilayer – Multilingual

Prospects and Challenges

Experiences: treebanks, grammars and languages

Case study TIGER: manifold uses for different types of grammars

ULA for multi-layer multi-lingual corpus annotation

Zeroing out or bridging differences?

Case study SALSA: logical formalisation of multi-layer annotations

Moving on

Networks of Linguistic Annotation – the Linguist's Web

Treebanks: varieties and usages

Varieties and dimensions of TB annotation

Monolingual and monostratal

Scattered across frameworks and languages

Multi-layered – Multi-lingual – Grammar induction

Monolingual extensions (cf. Penn-II TB, PDB, TIGER/SALSA, ...)

Multilingual treebanks – multilingual resource/grammar induction

Issues: standardisation and integration

Usages

Benchmarking – Resource Induction

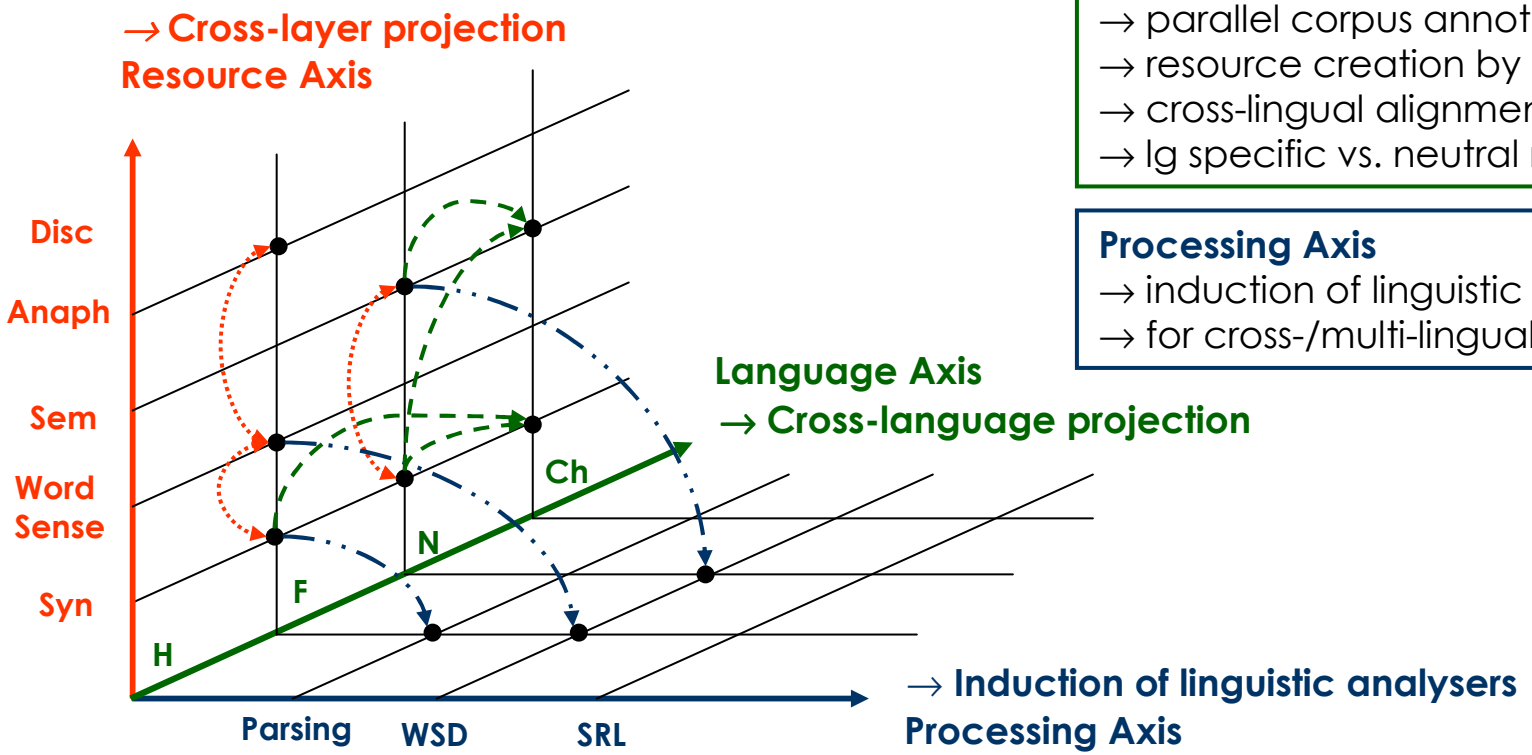
Monolingual vs. Multilingual NLP applications

Multi-layer multi-lingual annotations

Resource Axis
→ corpus annotation standards
→ multiple levels of linguistic annotation

Language Axis
→ parallel corpus annotation
→ resource creation by projection
→ cross-lingual alignment at many levels
→ lg specific vs. neutral representation

Processing Axis
→ induction of linguistic analysers
→ for cross-/multi-lingual applications



Resource Axis

Resource Axis: from mono-stratal to multi-layer annotation

Stand-off annotation

Multi-layer annotation schemes

Cross-referencing between layers

Prospects: Interaction across layers

Multi-layered grammar architectures

Multi-layer constraints for grammar induction and disambiguation

Limits and Challenges

Divergences of frameworks

Conflicting annotations

Interoperability

Language Axis

Using multilingual parallel /comparable corpora we could

Speed up annotation

Using cross-language projection techniques

Provide richer annotations

Exploiting multi-layer constraints via multi-level cross-lingual alignments

Harmonise differences across languages

Maximally uniform annotations qua projection

Studying and accommodating for divergences

Limits and Challenges

Noisy annotations

Translational divergence

Direct correspondence assumption does not hold uniformly

Processing Axis

Induction of multi-lingual automatic analysers

For different levels of analysis using multi-layer constraints

In shorter time

Using annotation projection and robust induction techniques

Limits and Challenges

Accessing different layers

Quality of multi-layer multi-lingual annotation

Prospects and Challenges

Prospects

Obtain multi-layer annotated corpora that are ***worth more than their sum***

For broad-coverage, efficient multilingual resource induction

Harmonisation across languages for multilingual applications

Challenges

from the viewpoint of ***corpus-based grammar induction and evaluation***

that may be useful for designing multi-layer multi-lingual annotation

- **Diversity of frameworks**
  - Exploitation of treebanks for grammar induction and evaluation
- **A Grammar's/Treebank's Lifetime**
  - Dynamic treebanks, moving treebanks (cf. Oepen et al. 2002)
- **Cross-level and cross-lingual referencing**
  - Cross-referencing/accessing potentially conflicting layers of representation

TB-based grammar induction and evaluation

Case study:

NEGRA/TIGER treebank

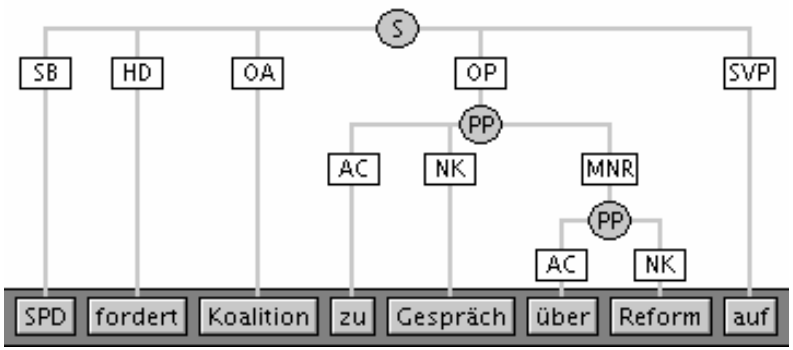
- Constituency and dependency annotation, enriched with morphology

TIGER dependency bank

- „Theory-neutral“ dependency representation

Heavily exploited for grammar induction

„treebank conversion“: converting TB to framework-specific structures




```
case(Museum~1, nom),
cmpd_form(Museum~1, Privatmuseum),
gend(Museum~1, neut),
mod(Museum~1, privat~1001),
mood(müssen~0, indicative),
num(Museum~1, sg),
oc_inf(müssen~0, weichen~3),
pers(Museum~1, 3),
sb(müssen~0, Museum~1),
sb(weichen~3, Museum~1),
tense(müssen~0, pres)
```

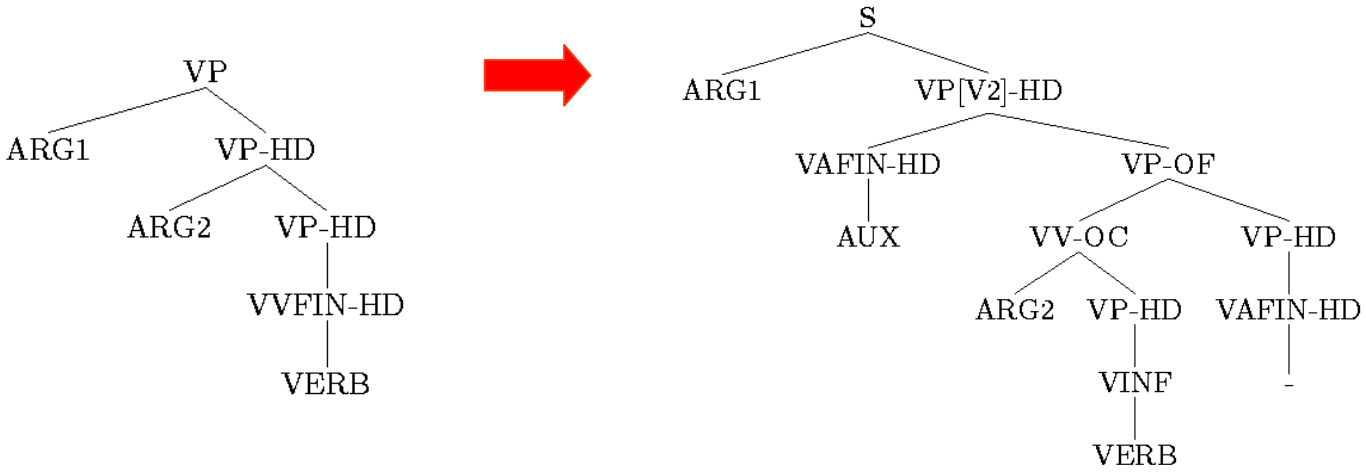
TB-based grammar induction and evaluation

Treebank conversion for LTAG grammar extraction (Frank 2002)

Binarisation of tree structure (tree adjunction)

Special phenomena (word order, coordination, extraposition, ...)

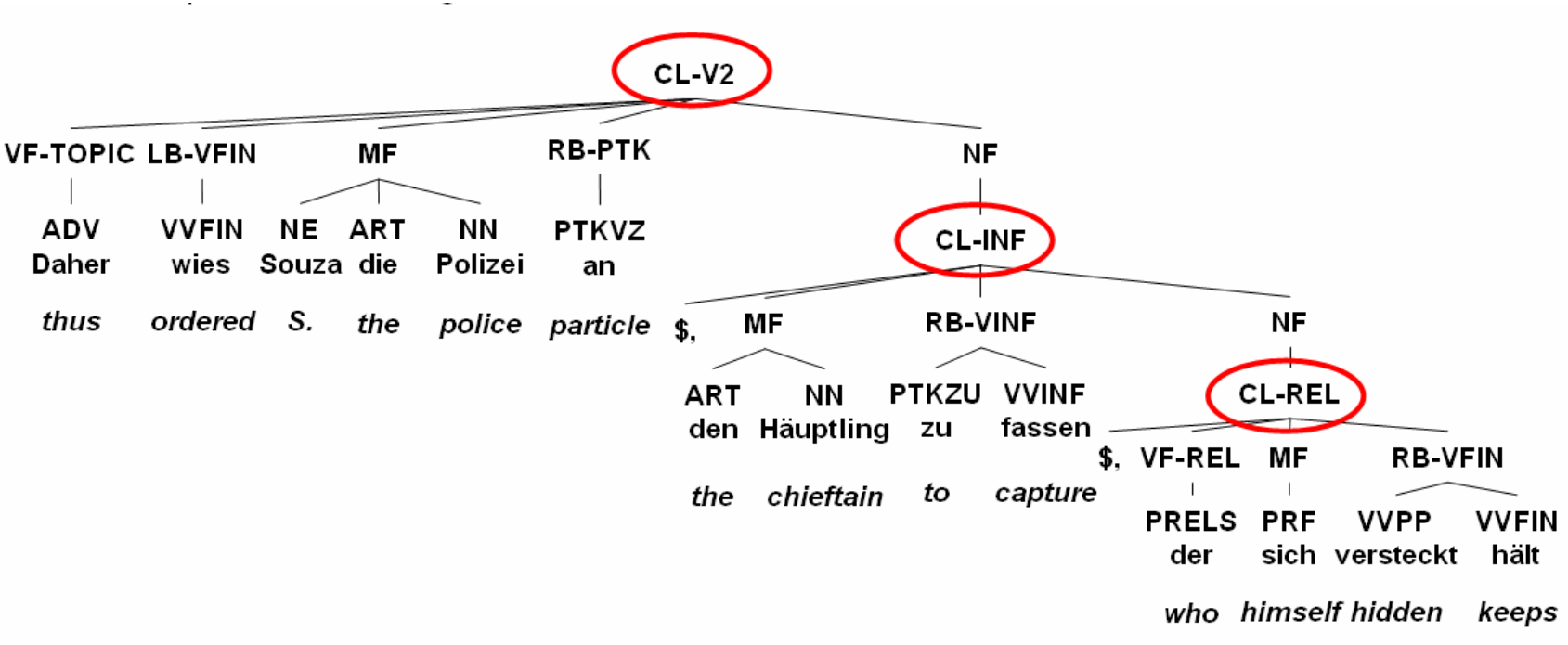
LTAG etree (subcat) extraction and tree family induction



(HPSG) subcat lexicon extraction

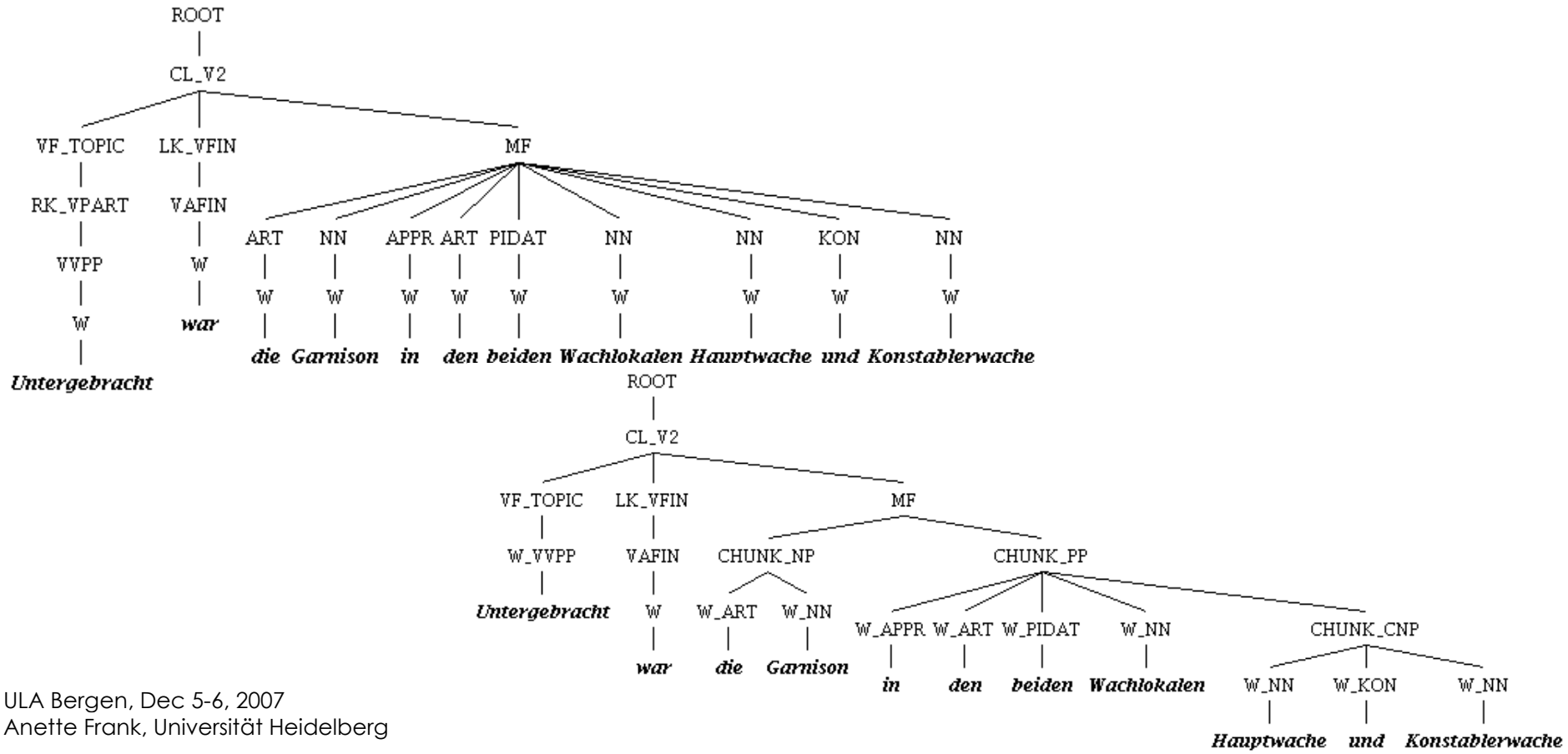
TB-based grammar induction and evaluation

Conversion to topological structures and parser induction (Becker and Frank 2002)



TB-based grammar induction and evaluation

Conversion to topological structures and parser induction (Becker and Frank 2002)



TB-based grammar induction and evaluation

Conversion to theory-neutral TIGER dependency bank (Forst et al, 2004)

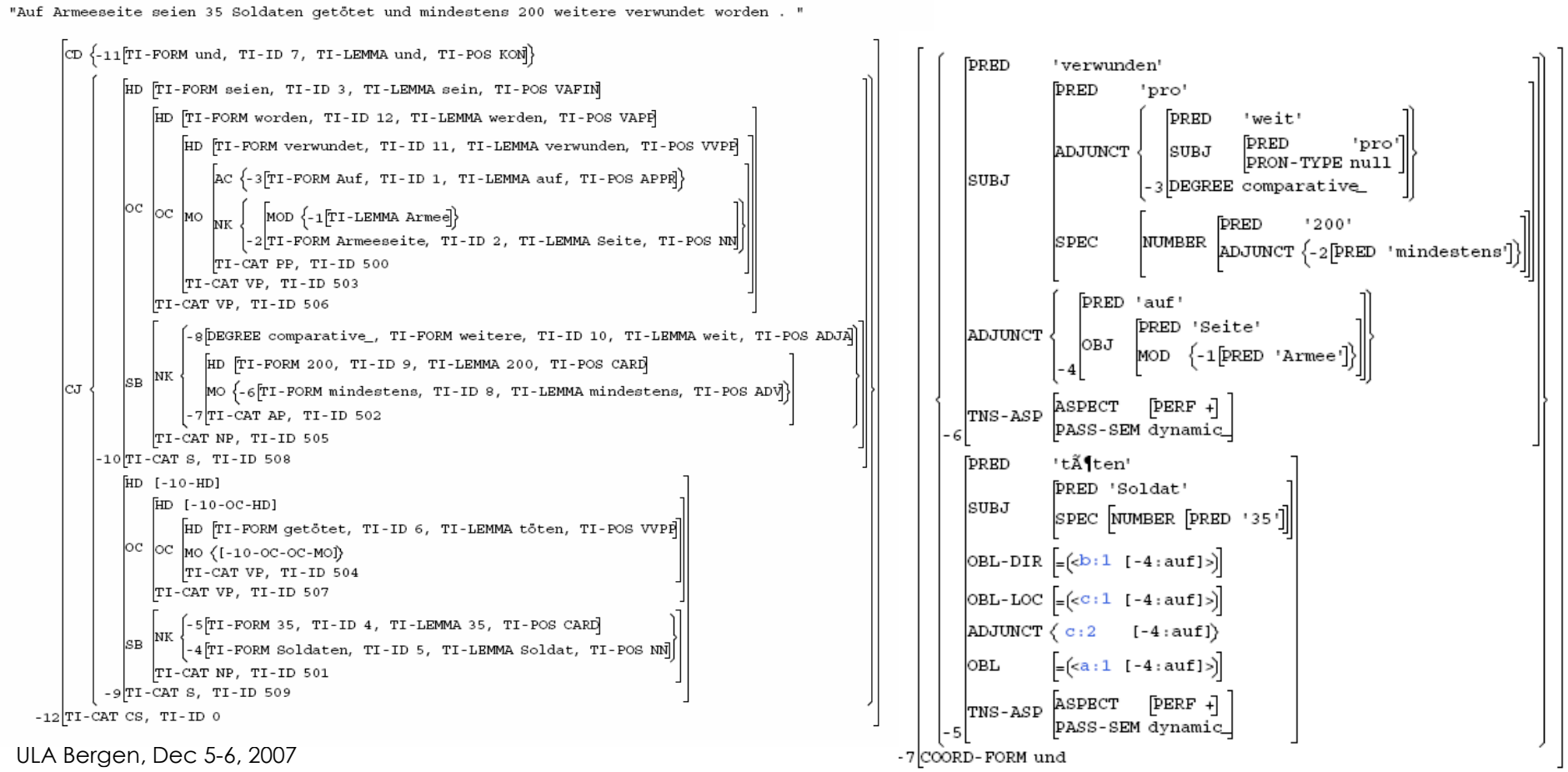
```
<s id="s8595">
 <graph root="s8595_500">
  <terminals>
   <t id="s8595_1"
     word="Privatmuseum"
     pos="NN" morph="Nom.Sg.Neut"/>
   <t id="s8595_2" word="muß"
     pos="VMFIN" morph="3.Sg.Pres.Ind"/>
   <t id="s8595_3" word="weichen"
     pos="VVINF" morph="--" />
  </terminals>
  <nonterminals>
   <nt id="s8595_500" cat="S">
    <edge label="SB" idref="s8595_1"/>
    <edge label="HD" idref="s8595_2"/>
    <edge label="OC" idref="s8595_3"/>
   </nt>
  </nonterminals>
 </graph>
</s>
```

```
case(Museum~1, nom),
cmpd_form(Museum~1, Privatmuseum),
gend(Museum~1, neut),
mod(Museum~1, privat~1001),
mood(müssen~0, indicative),
num(Museum~1, sg),
oc_inf(müssen~0, weichen~3),
pers(Museum~1, 3),
sb(müssen~0, Museum~1),
sb(weichen~3, Museum~1),
tense(müssen~0, pres)
```

TB-based grammar induction and evaluation

Conversion to f-structure treebank (Forst 2003)

Parser evaluation and disambiguation



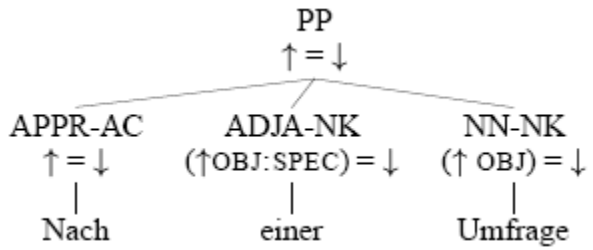
TB-based grammar induction and evaluation

Induction of LFG Grammar (Cahill et al. 2003)

Dealing with flat structures

```
(TOP
  ($*LRB* '')
  (S[up=down]
    (CNP-SB[up-subj=down]
      (NN-CJ[down-elem=up:conj]
                        Geschäftemachen)
      (*T1*-CD -)
      (*T2*-CJ -)
    )
    (VAFIN-HD[up=down] ist)
    (NP-PD[up-xcomp_pred=down]
      (PPOSAT-NK[up-spec:poss=down] seine)
      (NN-NK[up=down] Welt)
    )
    (KON-*T1*[up:subj=down] und)
    (NP-*T2*[down-elem=up:subj:conj]
      (PTKNEG-NG[down-elem=up:adjunct]
                             nicht)
      (ART-NK[up-spec:det=down] die)
      (NN-NK[up=down] Politik)
    )
  )
  ($. .)
)
```

```
subj : conj : 1 : pred : 'Geschäftemachen'
               2 : spec : det : pred : die
                   adjunct : 3 : pred : nicht
                   pred : 'Politik'
       coord_form : und
xcomp_pred : spec : poss : pred : pro
             pred : 'Welt'
pred : ist
```

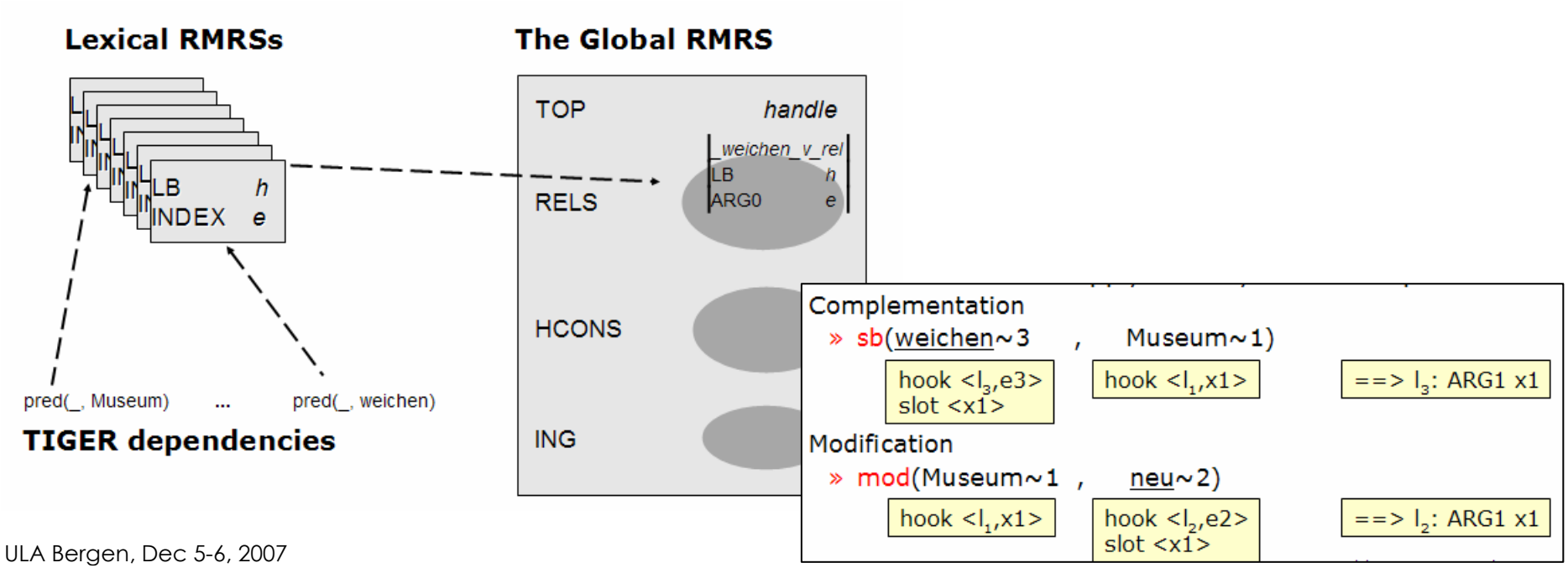


TB-based grammar induction and evaluation

Conversion to RMRS treebank (Spreyer and Frank 2005)

RMRS Semantics from TIGER DepBank

HPSG grammar evaluation

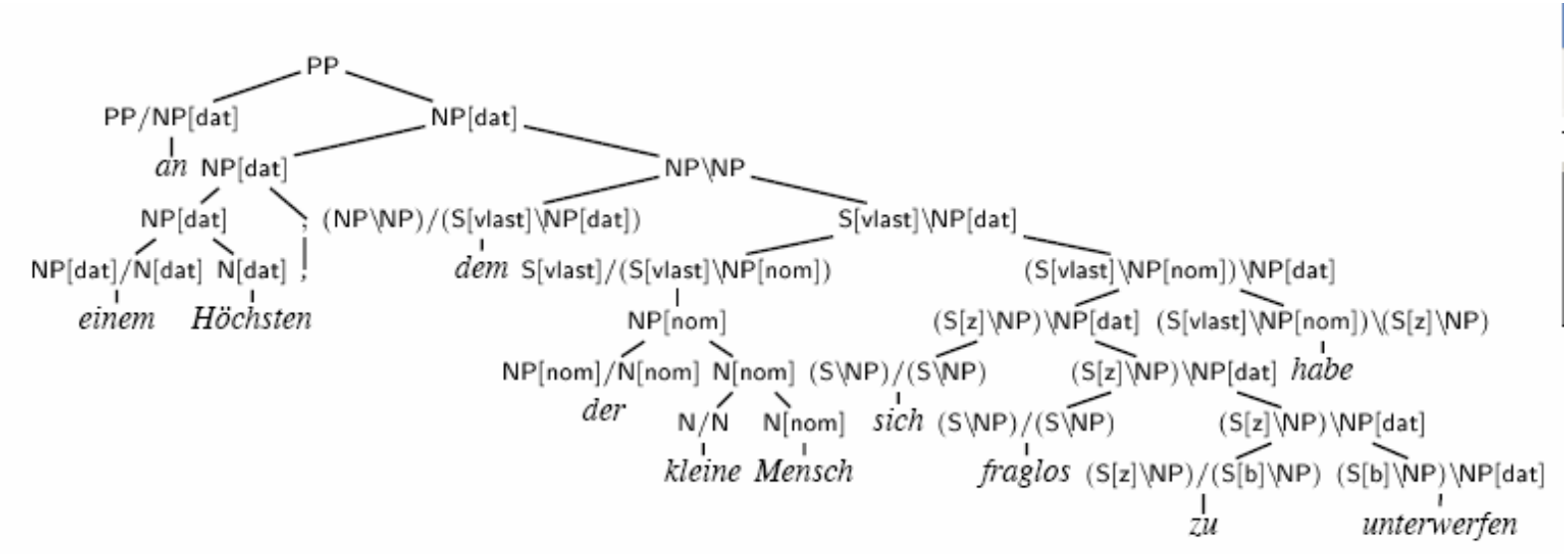


TB-based grammar induction and evaluation

Conversion to CCG (Hockenmaier 2006)

Binarisation of tree structures

Handling special phenomena (coordination, ...)



Lexicon extraction ≈ CCG grammar induction

TB-based grammar induction and evaluation

Grammar induction

Dubey and Keller: original TIGER annotation format

Frank and Becker: topological sentence structure

Cahill et al: LFG grammar induction

Hockenmaier: CCG grammar induction

Benchmarking and Evaluation

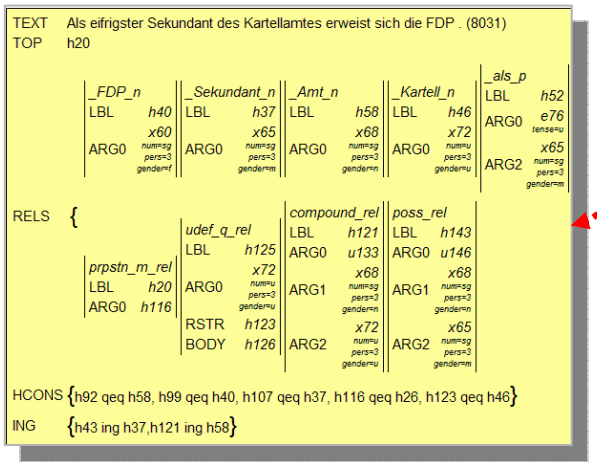
Forst: LFG grammar evaluation and disambiguation

Spreyer and Frank: HPSG grammar evaluation

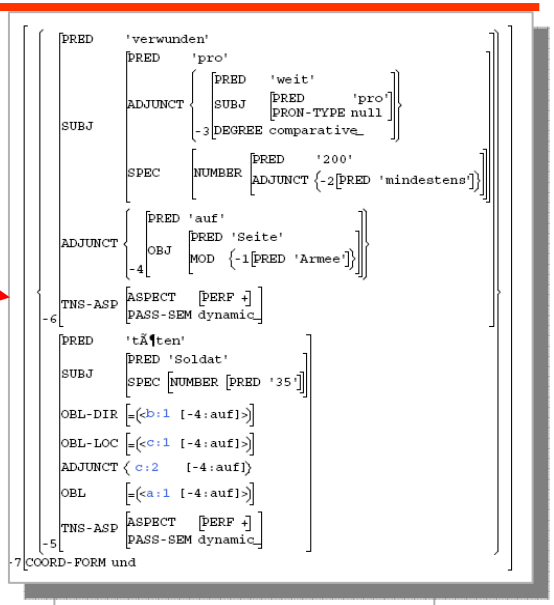
TB-based grammar induction and evaluation

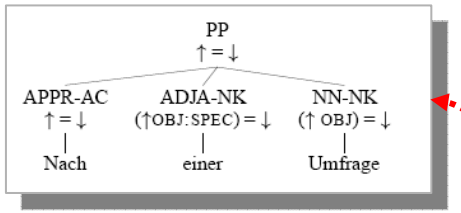
dependency-based mappings

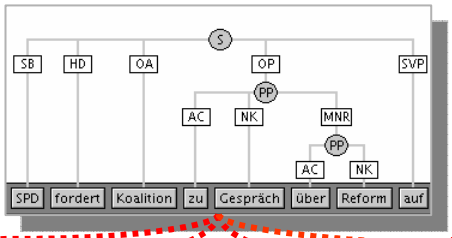
framework-specific structure mappings

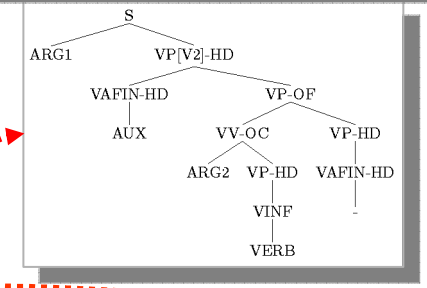


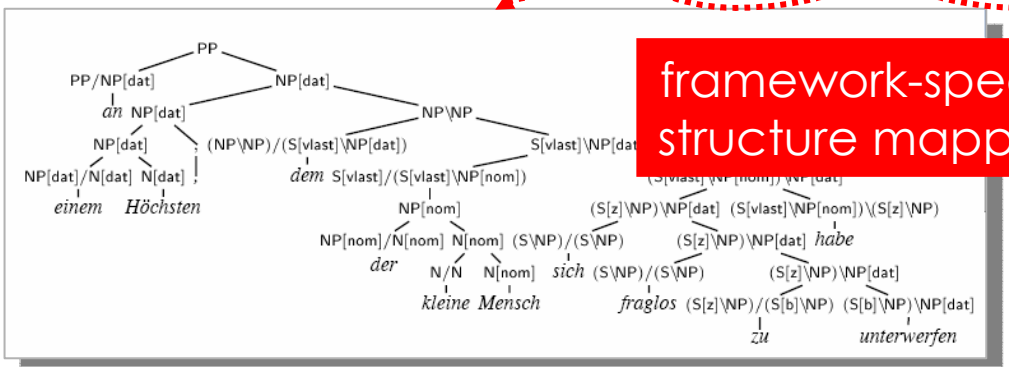

```
case(Museum~1, nom),
cmpd_form(Museum~1, Privatmuseum),
gend(Museum~1, neut),
mod(Museum~1, privat~1001),
mood(müssen~0, indicative),
num(Museum~1, sg),
oc_inf(müssen~0, weichen~3),
pers(Museum~1, 3),
sb(müssen~0, Museum~1),
sb(weichen~3, Museum~1),
tense(müssen~0, pres)
```

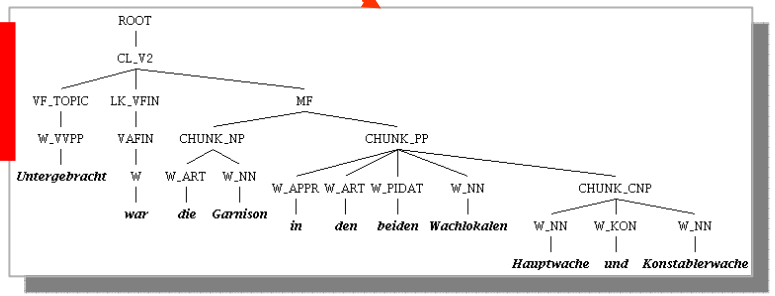












Lessons and reflections

What makes these frameworks so different

so that a single treebank is not enough?

Main Representation Levels

Syntax-Semantics Interface

Lexicalisation and Argument Structure

Views on Grammar Architecture

Main Representation Levels

Syntax-Semantics Interface

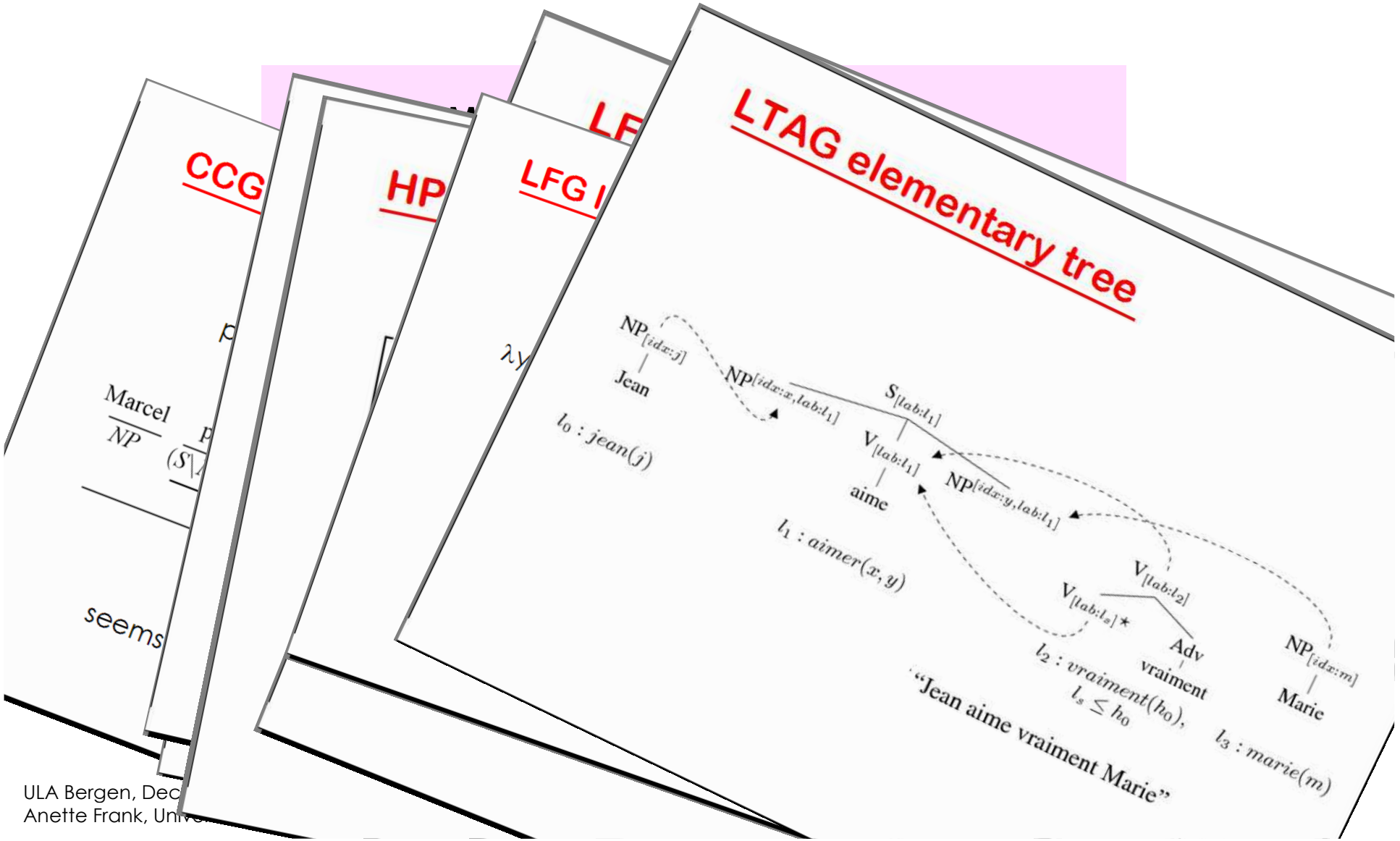
Semantic composition tightly joint with constituent structure

HPSG and (C)CG

Function-/dependency-driven semantic composition

LFG and L/FTAG

Main Representation Levels



Lessons and reflections

„Surface syntactic structure“

LFG, HPSG: can deal with „nonstandard“ phrase structure (extra cost)
LTAG, CCG: require framework-specific phrase structure

Treebank conversion

Mandatory for certain frameworks (LTAG, CCG)
Helpful for others (LFG f-structure induction, RMRS construction)

Mapping: graph-based rewrite rules (term rewriting)

Annotation / Treebank life time

Evaluation: sensitive to grammar changes

Induction: analyses dependent on original or converted TB

Monitoring grammar changes and *moving treebanks (term rewriting)*

Lessons and reflections

Diversity of frameworks

Structure mappings

Principled correspondences – idiosyncrasies

Lexicon and Argument structure

more direct correspondences across frameworks:

etrees / grammatical functions / subcat list / complex categories

Semantics

Lexicalisation, Scope and Underspecification

Variety of formalisms (UDRT, CLLS, RMRS) and equivalence results

Glue Semantics: LFG/LTAG/HPSG → (R)MRS, (U)DRT, λ-Calculus

Diversity and m-layer m-lingual annotation

Unified Linguistic Annotation?

Dealing with diversity

„Neutral“ annotation w/ mappings to framework-specific representations
Modelling ***principled*** correspondences via ***partial mappings***
Singling out special divergences for intricate phenomena

Standardisation

Across frameworks – Across languages
By way of ***(partial) correspondences***

Multi-layer multi-lingual corpus annotation

Zeroing out, merging or bridging differences?
Case study SALSA: logical formalisation of multi-layer annotations

Modeling divergences in multi-layer corpora

Case study SALSA: logical formalisation of multi-layer annotations

Annotation of FrameNet frames on top of TIGER syntactic structure

Multi-layer annotation

Aims

Corpus *and* Lexical Resource

Generic syntax-semantics interface

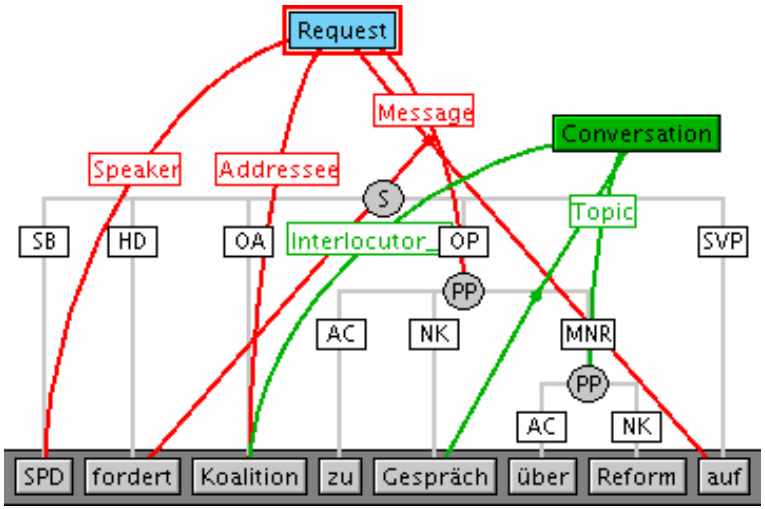
Issues

Consistency in annotation samples

Normalising granularity differences

Logics-based formalisation of

- multi-layer corpus annotations
- a lexicon model



Encoded in TIGER/SALSA XML: modular description of syntax and (frame) semantics

A Description Logics based Lexicon Model

Work by Dennis Spohr, IMS Stuttgart and SALSA (Spohr et al 2006/07)

From Corpus to Lexicon

Extracting a *frame-based lexicon* with *role-linking information* from corpus

Abstracting lexicon data from annotation samples

- Consistency checks
- Frequency information and data distribution
- Generalising fine-grained categories for combination with rule-based systems

DL-based modelling of FrameNet data (in OWL DL)

Formalisation of definitional part of FrameNet and corpus annotations

OWL DL

- Monotonicity, decidability (Baader et al. 2003)
- Reasoning and consistency checking services
- Storage and querying architecture (SESAME and SeRQL)

A Description Logics based Lexicon Model

T-Box

Linguistic Model

- FrameNet
  - Frames, Frame Relations
  - Roles
- Sense Assignment
  - Lemma – Frame
- Role Assignment
  - Syntactic units – Roles

Annotation Model

- Annotation Types
  - Frames: single, elliptic, metaphoric, USP
  - Roles: Single, USP
  - Target: Single, Multi-Word
- Sentences
- Syntactic units

Normalisation
Querying
Consistency checking

A-Box

Corpus: Annotation instances

- Sentences
- Syntactic units
- Frame and role annotations

T-Box vs. A-Box

T-Box: Linguistic and Annotation Model

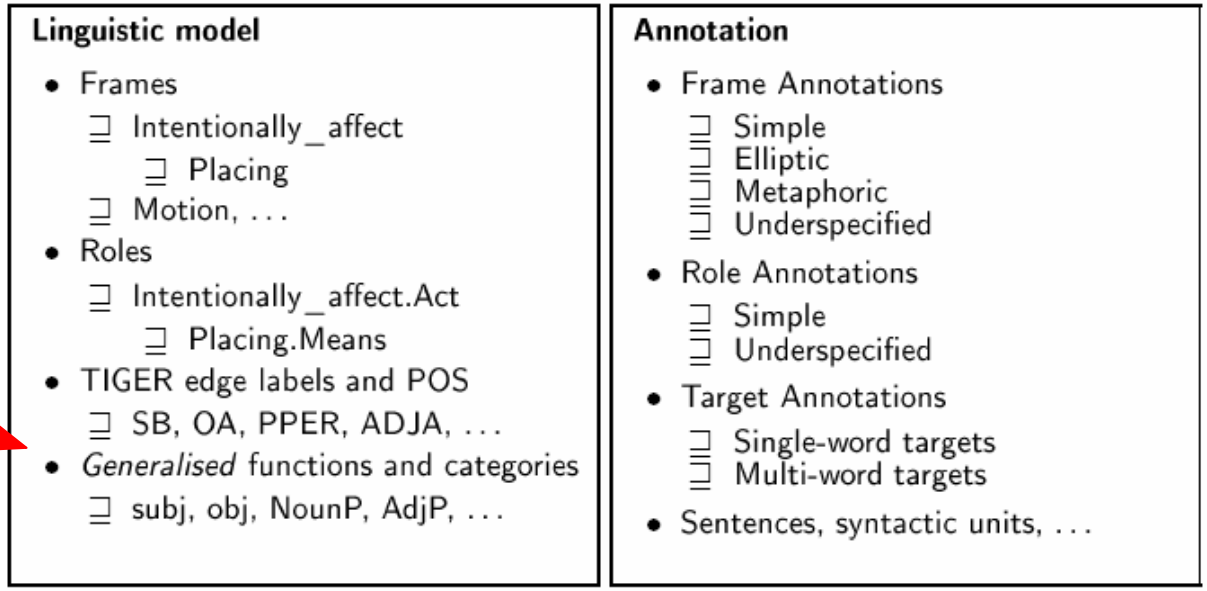
General ***and*** specific frame classes

General (normalised) ***and*** specific syntactic classes and functions

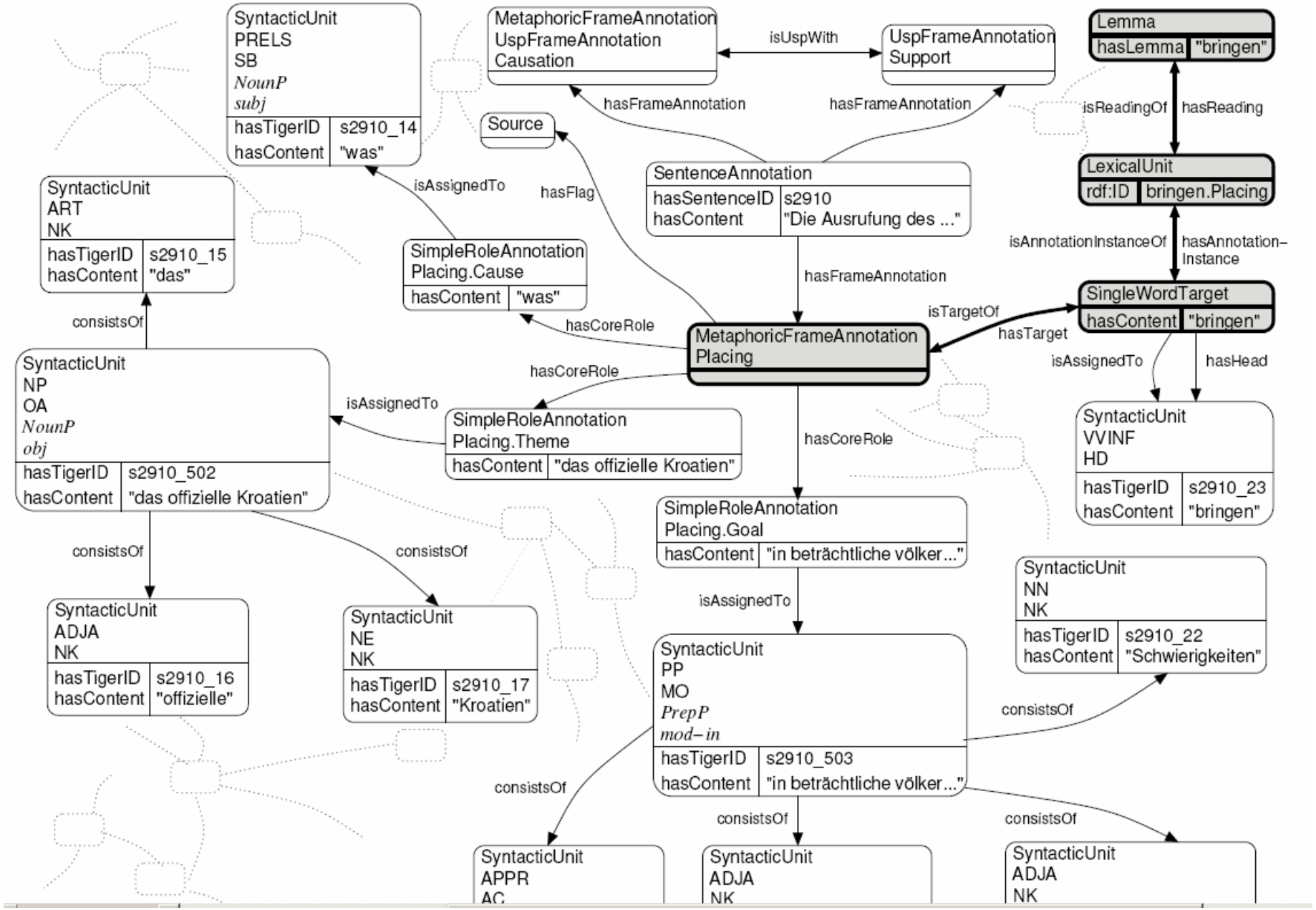
Defining "views" over annotated data: for modelling and retrieving

A-Box: corpus annotations

Consistency checking **through model checking**



Defining normalising categories


SyntacticUnit
PRELS
SB
NounP
subj
hasTigerID s2910_14
hasContent "was"
MetaphoricFrameAnnotation
UspFrameAnnotation
Causation
isUspWith
UspFrameAnnotation
Support
Lemma
hasLemma "bringen"
Source
hasFrameAnnotation
hasFrameAnnotation
isReadingOf
hasReading
SentenceAnnotation
hasSentenceID s2910
hasContent "Die Ausrufung des ..."
LexicalUnit
rdf:ID bringen.Placing
SyntacticUnit
ART
NK
hasTigerID s2910_15
hasContent "das"
isAssignedTo
hasFlag
isAnnotationInstanceOf
hasAnnotation–Instance
SimpleRoleAnnotation
Placing.Cause
hasContent "was"
hasFrameAnnotation
SingleWordTarget
hasContent "bringen"
consistsOf
hasCoreRole
isTargetOf
hasTarget
MetaphoricFrameAnnotation
Placing
isAssignedTo
hasHead
hasCoreRole
SyntacticUnit
NP
OA
NounP
obj
hasTigerID s2910_502
hasContent "das offizielle Kroatien"
isAssignedTo
SimpleRoleAnnotation
Placing.Theme
hasContent "das offizielle Kroatien"
hasCoreRole
SyntacticUnit
VVINF
HD
hasTigerID s2910_23
hasContent "bringen"
SimpleRoleAnnotation
Placing.Goal
hasContent "in beträchtliche völker..."
consistsOf
consistsOf
SyntacticUnit
NN
NK
hasTigerID s2910_22
hasContent "Schwierigkeiten"
SyntacticUnit
ADJA
NK
hasTigerID s2910_16
hasContent "offizielle"
SyntacticUnit
NE
NK
hasTigerID s2910_17
hasContent "Kroatien"
isAssignedTo
SyntacticUnit
PP
MO
PrepP
mod–in
hasTigerID s2910_503
hasContent "in beträchtliche völker..."
consistsOf
consistsOf
consistsOf
consistsOf
consistsOf
SyntacticUnit
APPR
AC
SyntacticUnit
ADJA
NK
SyntacticUnit
ADJA
NK

Querying: retrieving, counting, grouping

Retrieving information from the Corpus/Lexicon

- Queries specify paths through the model graph
- Allow querying of intersecting hierarchies
- Frequency counts, grouping and filtering data

Example: Extract *all lemmas that evoke the PLACING frame*

- Retrieved information (with grouping for frequency information)

Lemma	*No. of instances*	*Lemma*	*No. of instances*
legen	38	ablegen	3
bringen	35	kippen	3
nehmen	13	einlagern	1
plazieren	4	einpflanzen	1

Normalisation

Normalisation of linguistic information at different levels

- TIGER syntactic categories and edge labels
- Normalised syntactic categories and grammatical functions
  - NounP, PrepP, Sent, …. Subj, Obj, Pobj, …

Example: syntactic realisation of semantic roles

- Specific categories: 2.176 realisation patterns
- Normalised categories: 1.026 realisation patterns

Specific		Normalised	
Role	*Category/POS*	*Role*	*Category*
Placing.Theme	NN	Placing.Theme	NounP
Placing.Theme	NE	–	
Placing.Theme	PPER	–	
Statement.Message	S	Statement.Message	Sent
Statement.Message	VROOT	–	

Networks of Linguistic Annotation

Unified Linguistic Annotation or Bridging Differences?

Modelling *principled* correspondences via *partial mappings*

Defining interfaces – correspondences – partial mappings
between alternative „views“ on the data, at suitable abstraction level

Examples

local arguments, nonlocal dependency paths
modifiers, determiners, adjuncts, special constructions (MWE, support, ...)
role linking information, ..

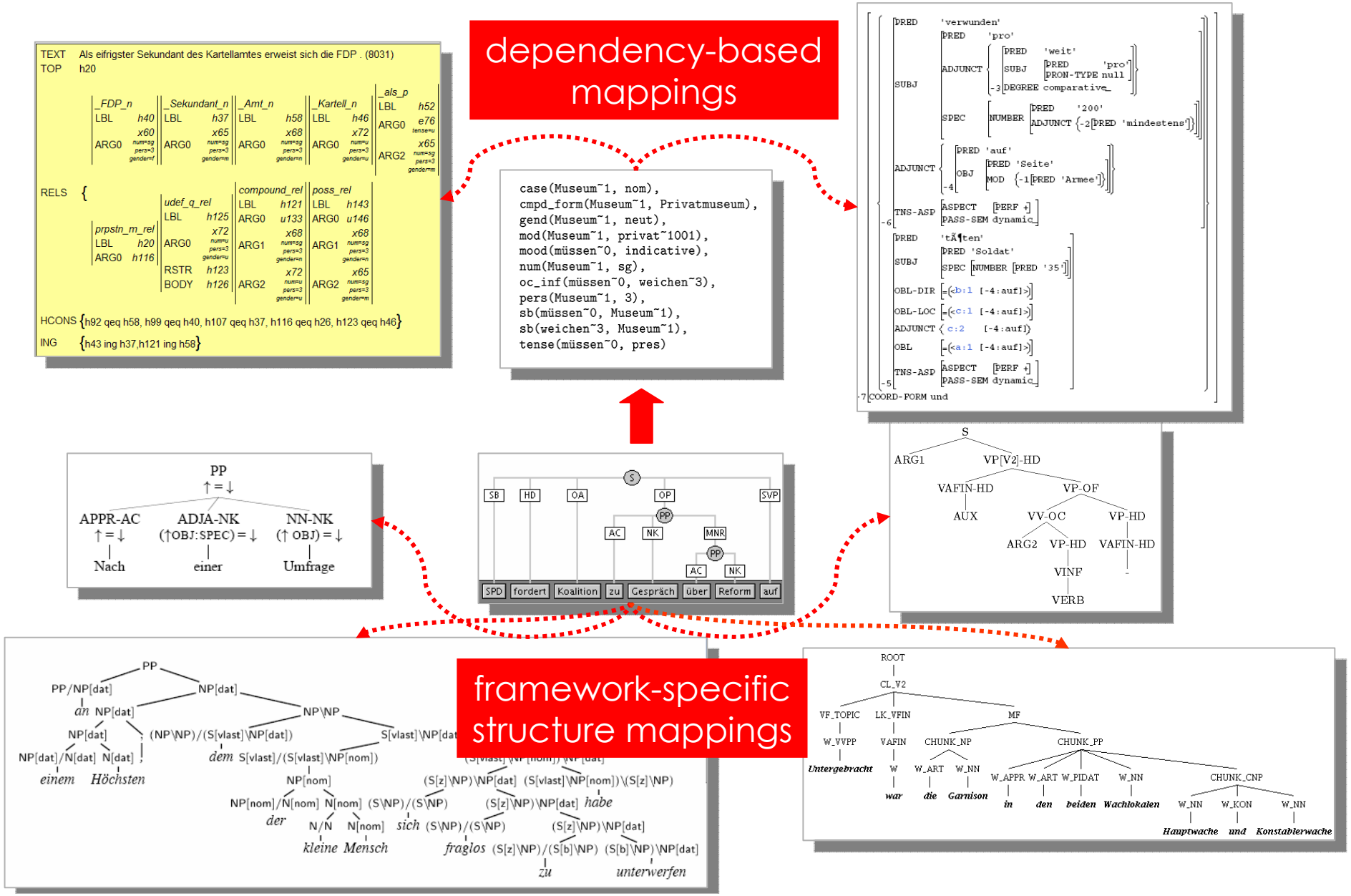
Translating „mappings“ to „correspondence“ relations

Relating framework-specific „configurations“ via „corresponds-to“
Singling out special divergences for intricate phenomena

Defining correspondences across languages

By way of ***(partial) correspondences***

From complex mappings to partial correspondences



Modelling multi-layer multi-lingual corpora

Breaking down complex mappings into (partial) correspondences

- Logical formalism: define complex constraints for correspondences
- Framework-specific annotations can remain stable
- Annotators can concentrate on „their theory“

Building a network of linguistic correspondences: the Linguist‘s Web

- Across linguistic layers
- Across languages

Defining and Refining

- Flexibly refine „views“ on the multi-layered multi-lingual corpus, according to evolving knowledge, needs and applications
- Start with „agreed“ and „general“ correspondences
- Extend towards intricate cases

Thank you