A tool for Corpus Analysis using partial Disambiguation and Bootstrapping of the Lexicon

Kurt Eberle, Ulrich Heid, Manuel Kountz, Kerstin Eckart

SFB-732, B3, Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
– Computerlinguistik –
Azenbergstr. 12
D 70174 Stuttgart

KONVENS 2008, Berlin

Overview

- Motivation and Objectives
- Objectives – State of the Art – Requirements
- Sample data and descriptive problems
- Tool Architecture – sample Representations
- Tool Functions and Applications
- Conclusions and future work

Motivation and Objectives

Ambiguity and data extraction from corpora

- Observation: structural ambiguity
  Er löst schwierige Aufgaben und Probleme
  - Task: extract verb+object pairs:
    * *Aufgabe* + *lösen*: ok
    * *Problem* + *lösen*: ok
    ⇒ No problem for extraction
  - Task: extract adjective+noun pairs:
    * *schwierig* + *Aufgabe*: ok
    * *schwierig* + *Problem*: ??
    ⇒ Structural ambiguity with impact on data extraction
- Objective 1:
  Ambiguity awareness in corpus exploration:
  not all ambiguities affect extraction

Motivation and Objectives

Data extraction: from targeted phenomenon to extracted data (1/2)

- Trivial observation:
  Data extraction involves several steps:
  - Knowledge of phenomenon to be extracted
  - Query formulation, application to (preprocessed) text
  - Query result, including local analysis
  - Interpretation
- Even on preprocessed corpora:
  Some phenomena are hard to express in terms of queries
- Objective 2:
  Make relation visible and reproducible between:
  - corpus sentence
  - analysis of preprocessing + query
  - parameters used for disambiguation

$\Rightarrow$ Integrated representation

Motivation and Objectives

Data extraction: from targeted phenomenon to extracted data (2/2)

- Observation:
  To test descriptive and extraction hypotheses:
  - Need to find analogous cases in corpus, with respect to
    * analysis of preprocessing
    * query
    * parameters used for disambiguation
- Objective 3:
  Integrated representation
  should be fully searchable:
  all aspects separately – together – any partial combination

Motivation and Objectives

Using data extraction to test descriptive hypotheses

- Observation:
  To test descriptive hypotheses, often
  a few contextual parameters are modified,
  all others being kept the same
- Objective 4:
  Keep track of variable depth of analysis
  (cf. Kay/Gawron/Norvig 1994):
  possibility to vary the degree of abstraction of analysis representations

Motivation and Objectives

Bootstrapping in corpus-based data extraction

- Observation:
  Data extraction is similar to hypothesis testing:
  improvement through results obtained in intermediate steps:
  search $\longrightarrow$ evaluation of results $\longrightarrow$ refined search $\longrightarrow$ better results
- Objective 5:
  Enable bootstrapping of data extraction
  by making intermediate results
  available to the search tool

Objectives vs. the state of the art

Could we use existing tools and resources?

	Objectives: reminder	Corpus tools and resources	Parsing
1	Ambiguity awareness		Deep parsing: formal grammars
2	Integrated representation: sentence/analysis/ disamb. parameters	treebanks, parsed coprora	Search in parse forests
3	Queriable integrated representation	treebank query	
4	Variable depth	multilayered corpora	grammar
5	Bootstrapping	machine learning?	Feedback to parser

Objectives vs. the state of the art

Using existing technology (1/2)

Corpus tools and resources

- Treebanks are typically disambiguated
- Size of treebanks: rather small
- Multilayered corpora:
  Possible in principle – tools and resources: prototypes
- Machine learning:
  Parameters not clear enough in many cases

Objectives vs. the state of the art

Using existing technology (2/2)

Parsing

- Deep parsing: appropriate;
  We prefer dependency-style grammars,
  where variable depth of analysis
  is easier to realize
- Search in parse forests is hard;
  We prefer underspecified representations
- Feedback to parser:
  via flexible lexicon interface

From objectives to requirements

Outline for an architecture

Integrated tool for analysis and data extraction:

(1) ambiguity awareness:
dependency-based parsing
and underspecified representation of analysis results

(2) integrated representation:
sentence – analysis/analyses – disambiguation parameters

(3) integrated representation should be searchable

(4) variable depth of analysis:
underspecification and partial disambiguation

(5) bootstrapping:
via flexible interface to the lexicon:
easy way to insert extraction results into lexicon

Sample data – descriptive problems

Sortal ambiguity of nominalizations as an extraction problem

German *-ung*-nominalizations are sortally ambiguous:

- *Teilung*: event (e) or result state (s)
- *Messung*: event (e) or object (data,o)
- *Abdeckung*: event (e), state (s) or object (o)

Sortal disambiguation depends on ...

- ... verb semantics (Roßdeutscher et al. 2007): e.g.
  - deadjectival, property-denoting: e,s (*trocknen, lähmen*)
  - effected object: e,o (*sammeln, bilden*)
  - denominal, affected object: e,s,o (*pflastern*)
- ... contextual clues

Sample data – descriptive problems

Contextual clues for sortal disambiguation: indicators

Context partners of nominalizations as possible sort indicators

- Modifiers, e.g.
  - Attributive adjectives: *kontinuierliche Messungen* (e)
  - Temporal/local PPs: *Messungen im Januar* (e)
  - Complement-indicating genitives: *die Fälschung von antiken Münzen* (e)
- Selectors, e.g.
  - Embedding verbs (whose objects the nominals are): *Spaltung überwinden* (s), *Messung durchführen* (e), *Fälschung verkaufen* (o)

⇒ Task for data extraction:
Identify sortal reading(s) by use of contextual data,
But: not all context partners are indicators:
wir brauchen neue Messungen

Sample data – descriptive problems

Structural ambiguity and its impact on extraction

Indicator status may depend on structural reading:

- *Messungen in Gomel*: PP-loc $\Rightarrow$ e-Indicator
- *Messungen des Radiologischen Instituts in Gomel*:
  PP-attachment ambiguous:
  (a) [Messungen [des R.I.$_{NP-gen}$][in G.$_{PP-loc}$]] $\Rightarrow$ e
  (b) [Messungen [des R.I. [in G.]$_{NP-gen}$]] $\Rightarrow$? $\Rightarrow$ no indicator
- *Peterson plante stundenweise Absperrungen mit verschiedensten Materialien*
  (a) *stundenweise Absperrungen* $\Rightarrow$ e
  (b) *Absperrungen mit verschiedenen Materialien* $\Rightarrow$ o
  stundenweise planen "absorbs" potential e-Indicator

Sample data – descriptive problems

Unexpected readings of *-ung*-nominals: facts

Examples:

- *die Verschärfung der Gesetze hat bewirkt, daß ...*
- *die Messungen zeigen/beweisen, daß ...*
- *die Überzeugung, wonach ...*
- Fact-readings:
  - "the fact that the laws have been made more restrictive ..."
  - "the fact that [x] has been measured ..."

⇒ Task for data extraction:
Find verbs which license fact-readings of *-ung*-nominals as subjects:
same as verbs taking subject clauses?

Tool Architecture

Requirements – Specifications

- Flexible interface to the lexicon
  - "Messungen `(e,o)` zeigen `subj: (f,o,e)`... "
    (clash: `e` & `f` / no clash: `o` & `o`)
- Underspecified representations
  - search for sentences which **can** show investigated phenomena:
  - "Messungen des Instituts in Gomel zeigen"
- Partial disambiguations
  - concentrate on considered phenomena
  - avoid irrelevant detailing of readings
  - tractability, efficiency

Tool Architecture

Realization – Implementation

- Adaptation of the MT tool *translate* (research version)
  - cooperation with the developer *Lingenio GmbH*
  - dependeny based unification grammar
    slot grammar (McCord 89/91)
  - underspecified semantic representation
    Flat underspecified discourse representation theory (FUDRT)
    (Eberle 97/04)
    (Extension of UDRT (Reyle 93) for broad coverage representations, implemented during SFB 340)
  - Large grammars, very large dictionaries, good performance
    – IBM's Logic based Machine Translation project (LMT)
    – Continuous exposition to market for 15 years
    – corresponding continuous improvement and extension

Tool Architecture

Semantic representations (FUDRSs) – tool functions

- Lexical representations are functions
  (to be evaluated gradually by context)
- Sentence representations are sets
  of DRSs and DRS-modifiers and modifiers of . . .
  and statements about order and type of application
- Carry out partial disambiguation and compute effects
- Inspect underspecified representations
- Use FUDRSs in search queries

Tool Architecture - Lexical Representation

The example of *Absperrung* (e,s,o)

Absperrung defines a mapping from DRFs in labeled DRSs (FUDRSs)
evaluated (partly/gradually)
depending on contextual information acting as a trigger.

absperrung(ref) ⇒ nsem_l

absperrung(e @ event) : $l_{e@event}$:

e (s, o)
absperren(e)
abgesperrt(s)
meets(e,s)
absperrung(s,o)

absperrung(s @ state) : $l_{s@state}$:

s (e, o)
abgesperrt(s)
absperren(e)
meets(e,s)
absperrung(s,o)

absperrung(x @ object) : $l_{x@object}$:

x
absperrung(x)

e @ event/s @ state... conditions in the sense of *wait*-statements:
wait((absperrung(SORT: event) →_))

Tool Architecture - Lexical Representation

Alternative lexical entries

Make use of derivation rules. . .

$$\underline{\text{absperrung}}(x) := l_{x@eso}:\text{deriv}(l_{e@event}:\underline{\text{absperren}}(e))$$

Tool Architecture - Sentence Semantics

Structured sets of partial representations

LKA-Mann Peterson plante stundenweise Absperrungen mit verschiedenen Materialien.

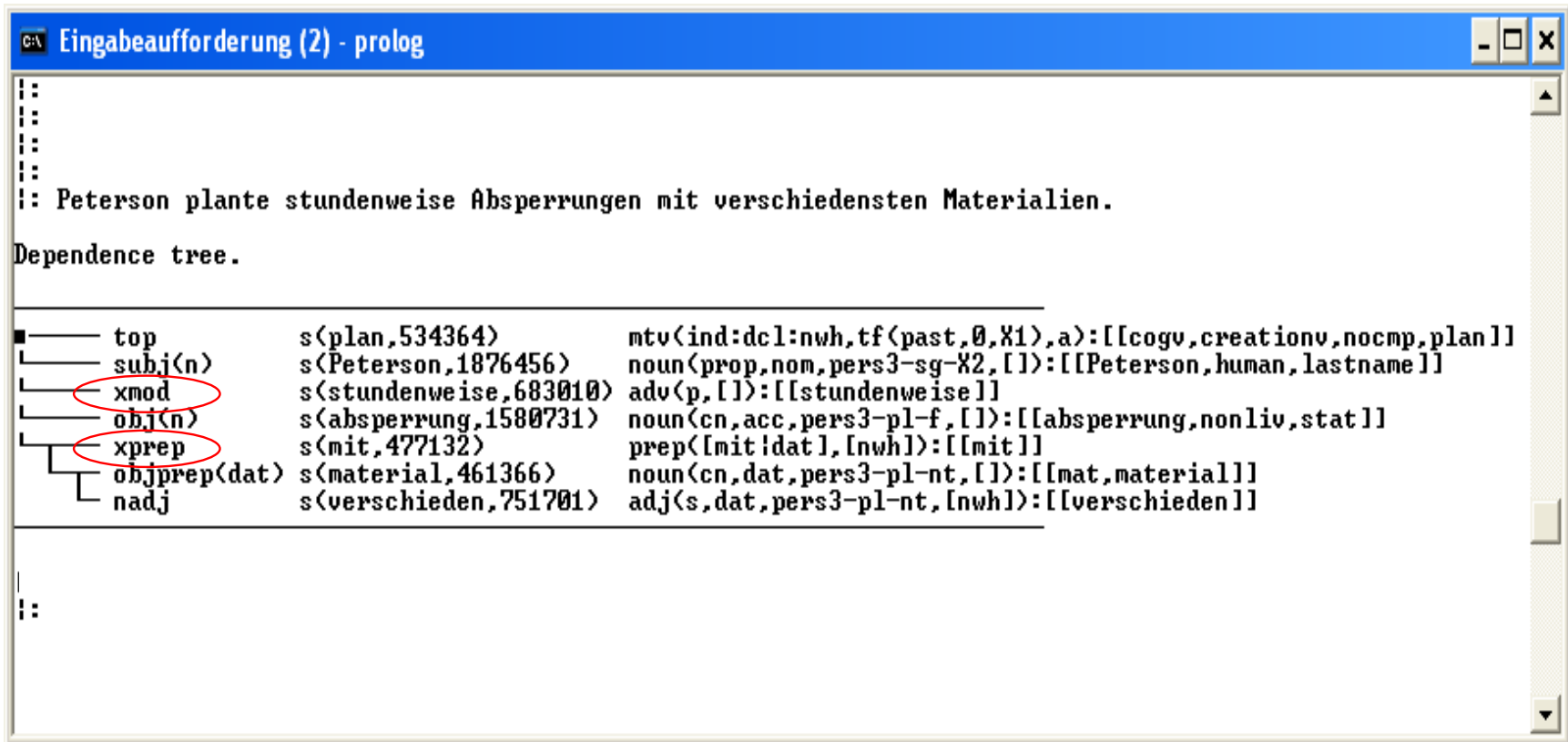
(i) Peterson planned hour-long blockings using all kinds of materials.

(ii) Repeatedly for hours, Peterson planned barriers consisting of all kinds of materials.

$$l0_e : \boxed{\begin{array}{l} e \\ \\ \text{planen}(e) \\ \text{agent}(e) = p \\ \text{theme}(e) = X \end{array}} \left\{ \begin{array}{l} l1_p : \boxed{\begin{array}{l} p \\ \text{peterson}(p) \\ \underline{\text{lka-mann}}(p) \end{array}}, \; l2_{E'} : \underline{\text{stundenweise}}(L'2_{e'@event}), l3_{X@eso} : \underline{\text{absperrung}}^*(X), \\ \\ l4_Y : \underline{\text{mit}}(l5_Y : \underline{\text{verschieden}}(\underline{\text{material}}^*(Y)), \; L5'_{z@object \vee e\_cog}) \end{array} \right\}$$

Tool Architecture - Sentence Semantics

Output of the *translate* implementation



Tool Architecture - Partial Disambiguation

Specifications: constrain interpretations

Peterson plante stundenweise Absperrungen mit verschiedensten Materialien.

(i) Peterson planned hour-long blockings, using all sorts of material.

(ii) Repeatedly for hours, Peterson planned barriers consisting of all sorts of material.

Specification (i)

$$\left\{ \begin{array}{l} l1_p : \boxed{\begin{array}{l} p \\ \text{peterson}(p) \\ \underline{\text{lka-mann}}(p) \end{array}},\ l2_{E'} : \underline{\text{stundenweise}}(L'2_{e'@event}), l3_{X@eso} : \underline{\text{absperrung}}^*(X), \\ \\ l4_Y : \underline{\text{mit}}(l5_Y : \underline{\text{verschieden}}(\underline{\text{material}}^*(Y)),\ L5'_{z@object \vee e\_cog}) \end{array} \right\}$$

$$l0_e : \boxed{\begin{array}{l} e \\ \\ \text{planen}(e) \\ \text{agent}(e) = p \\ \text{theme}(e) = X \end{array}} \quad \{\text{first}(l3,l2)\}$$

Tool Architecture - Partial Disambiguation

Specifications: consequences

Peterson plante stundenweise Absperrungen mit verschiedensten Materialien.

(i) Peterson planned hour-long blockings, using all sorts of material.

(ii) Repeatedly for hours, Peterson planned barriers consisting of all sorts of material.

Specification (i)

$$\left\{ \begin{array}{l} l1_p : \boxed{\begin{array}{l} p \\ \text{peterson}(p) \\ \underline{\text{lka-mann}}(p) \end{array}},\ l3_{E@event} : \underline{\text{absperrung}}^*(X) \left\{ \ l2_{E'} : \underline{\text{stundenweise}}(L'2_{e'@event}) \ \right\}, \\ \\ l4_Y : \underline{\text{mit}}(l5_Y : \underline{\text{verschieden}}(\underline{\text{material}}^*(Y)),\ L5'_{z@object \vee e\_cog}) \end{array} \right\}$$

$$l0_e : \boxed{\begin{array}{l} e \\ \\ \text{planen}(e) \\ \text{agent}(e) = p \\ \text{theme}(e) = X \end{array}}$$

Tool Architecture - Partial Disambiguation

Specifications: constraints and consequences

Peterson plante stundenweise Absperrungen mit verschiedensten Materialien.

(i) Peterson planned hour-long blockings, using all sorts of material.

(ii) Repeatedly for hours, Peterson planned barriers consisting of all sorts of material.

Specification (ii)

$$\left\{ \begin{array}{l} l1_p : \boxed{\begin{array}{l} p \\ \text{peterson}(p) \\ \underline{\text{lka-mann}}(p) \end{array}},\ l2_{E'} : \underline{\text{stundenweise}}(L'2_{e'@event}), l3_{X@object} : \underline{\text{absperrung}}^*(X), \\ \\ l4_Y : \underline{\text{mit}}(l5_Y : \underline{\text{verschieden}}(\underline{\text{material}}^*(Y)),\ L5'_{z@object \vee e\_cog}) \end{array} \right\}$$

$$l0_e : \boxed{\begin{array}{l} e \\ \\ \text{planen}(e) \\ \text{agent}(e) = p \\ \text{theme}(e) = X \end{array}} \quad \{\text{first}(l3, l4)\}$$

Tool Architecture - Partial Disambiguation

Constraints and consequences: tool output

Tool Functions and Applications

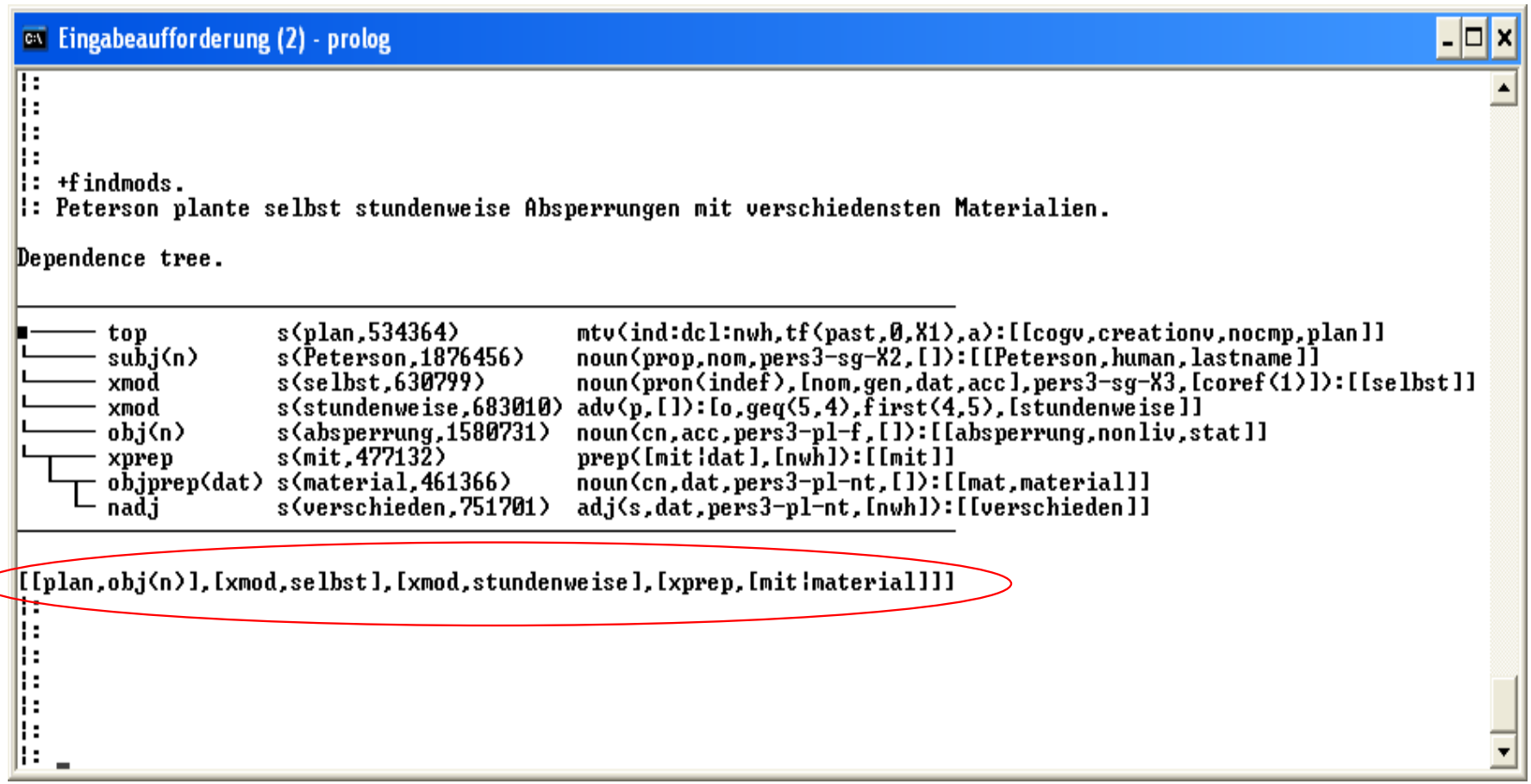
- Trigger partial disambiguation and compute effects
- Query representations
- Use underspecified representations in search queries

Tool Functions and Applications

Querying the integrated representation

Searching for modifiers and selectors of *-ung*-nominalizations

Tool output



Tool Functions and Applications

Searching for modifiers and selectors

Flexible output representations

Variable granularity: with/without morphosyntactic, semantic and contextual information...
without...

Selectors:
[feststell,vprep(bei),xprep(bei),xprep(nach)]
[gelt,iobj(p([bei|dat])),subj(n),xprep(von)]
...

Modifiers:
[nprep(in),frühjahr]
[xprep(in),grundwasser]
[nprep(in),monat]
...
[nadj,monatelang]
[xmod,stundenweise]
...

Tool Functions and Applications

Searching for modifiers and selectors – Flexible output representations

With morphosyntactic, semantic and contextual information...

[n(feststell,mtv(ind:dcl:nwh,tf(past,0,_),a), [cogv]), [xprep(bei), [166]], [xprep(nach), [293]]]
[n(feststell,mtv(ind:dcl:nwh,tf(past,0,_),p), [cogv]), [vprep(bei), [40,297]]]
...
[nprep(in), [n('frühjahr',det(def,sg), [season0]), [40]]]
[xprep(in), [n(grundwasser,det(def,sg), [liq]), [256]]]
[nprep(in), [n(monat,det(def,pl), [intervall0,timeas]), [257]]]
...

- identify regularities from syntactico-semantic information
  Example: [`xprep(in),intervall0`] as event indicator (?)
- incorporate indicator property into lexicon
- test assumption
- revise / adopt
- bootstrapping of the lexicon

Tool Functions and Applications

Querying underspecified representations – Search with unterspecified representations

Example: potential fact interpretations of *-ung*-nominalizations
Scenario:

- Hypothesis: subject of *zeigen* is interpreted as fact, if its representation is of type e (*event*)
- Test hypothesis:
  - Extract *zeigen*-sentences where subject **can** be *event*
  - Reject *zeigen*-sentences where it cannot or where *o* reading of *-ung*-subject seems preferred (because of missing additional hints)
    - "Messungen des Instituts . . . von . . . in Gomel. . ." can be *e*
    - "Messungen des Instituts in Gomel, die dort 99 veröffentlicht wurden,. . ." cannot
    - "Messungen des Instituts in Gomel, das erfahrener ist als das in Celle,. . ." gives no hint
- use corresponding underspecified representation in query!

Tool Functions and Applications

Querying underspecified representations – Search with unterspecified representations

Example: potential fact interpretations of -*ung*-nominalizations
Case: Sentences whose VP possibly characterizes facts and whose subj

- -*ung*-nominalization derived from e-root
  (has e-reading, Rossdeutscher 2007)
- subject of VP which may characterize facts
  (described by potential sentential complement)
- e-reading of nominalization possibly supported by additional e–modifier

FUDRS representation:

$$\{\ l1_{u@eso}\text{:ung\_deriv}(l'1_{e'@event}\text{:}L'1),\quad l2_{u'}\text{:}\underline{\text{mod}}(L'2_{x@eso}),\quad l3_{e''}\text{:}\underline{\text{fin}}(e'')\ \}$$

$$\{\diamond\ \text{first}(l1,l2)\ \}$$

$$l0_e\text{:}L0\ \&\ \boxed{\begin{array}{l}e\\ \text{subj}(e,u)\end{array}}$$

Tool Functions and Applications

Querying underspecified representations – Search with unterspecified representations

Excerpt from results (newspaper corpus *Wortschatz* University of Leipzig);

Die Verschärfung der Korruptionsgesetze im Jahr 1997 hat bewirkt, dass selbst das so genannte "Anfüttern" von Beamten mit Geschenken ohne Gegenleistung als strafbar gilt.
Und weiter heißt es in der vom braunen Zeitgeist geprägten Broschüre: "Dass die nationalsozialistische Gedankenwelt Allgemeingut der ganzen Kolonie geworden ist, zeigt die Beteiligung aller Volksgenossen am Winterhilfswerk und an gemeinsamen Eintopfessen."
Der Arbeitsalltag der Hausdame ist bestimmt von der gelebten Überzeugung, wonach eine Frau in ihrem Fach entweder "mit einem Mann oder einem Hotel verheiratet ist".

- potential e modification: *im Jahr 1997, am Winterhilfswerk*
- potential sentential complement *wonach . . .*
- 'deep structure' information *bestimmt von der gelebten Überzeugung*

Conclusions

A tool for the syntactico-semantic analysis of corpus text

- generates and uses underspecified representation in corpus analysis and search
- based on slot grammar, FUDRT
- implemented as adaptation of the reserach version of *translate*
- provides functions for
  - partial disambiguation
  - extraction of potential sortal indicators
  - testing of sortal theories

Conclusions

A tool for the syntactico-semantic analysis of corpus text

- has modular, flexible architecture
  - can use lexical information of variable granularity
  - supports bootstrapping of the lexicon
- currently used to test theories about sortal properties of *-ung*-nominalizations:
  detailed analysis of cases of sortal paradoxes, fact readings, ...
  Basis: fictional text (6000 analyses), newspaper corpus (30000 analyses)
- extensible to the investigation of other phenomena

Future work

Phenomena, tools and documentation

- extend investigation to other phenomena:
  subcategorization behaviour and nominalization of German particle and prefix verbs,
  extraction of collocations, ...
- bootstrap the system by corresponding results
- compare the results to results of other tools (Schiehlen 2003), ...
- implement database for storing and relating
  - corpus data
  - analyses of different tools,
    ranging over different levels of linguistic description
  - metadata
- stepwise improvement of reliablitity of corpus analyses and of analysis tools