SFB 732: Project B3
Status as of 11/2008,
Perspectives for 2009/2010

Ulrich Heid, Kurt Eberle, Kerstin Eckart

Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
– Computerlinguistik –
Azenbergstr. 12
D 70174 Stuttgart

Blaubeuren, 11/2008:
Input to the general presentation of area B
Objectives of B3: Reminder
Ambiguity-awareness in data extraction from text corpora

- Phenomena: Sortal ambiguity of German -ung-nominalizations:
  Teilung: e|s, Schätzung, Messung: e|o, Abdeckung: e|s|o,
  also fact readings: die Verurteilung von x zeigt...

- Linguistic viewpoint:
  - Which context partners (potentially) serve as sort indicators?
  - Under which circumstances?

- Corpus linguistic viewpoint:
  - frequency/productivity of phenomena
  - problem: ambiguity:
    * how to count? – relevant readings, not sentences
    * full disambiguation versus partial disambiguation:
      → phenomenon oriented disambiguation

- Tool architectures: How can an automatic system
  - find possible indicators?
  - (appropriately) disambiguate -ung-nominalizations?
Objectives of B3 within the SFB
Ambiguity-awareness in data extraction from corpora

General problem of corpus-based linguistics
How to make data extraction ambiguity-aware?

- In terms of the representation of analyses:
  FUDRT as project-internal representation
- In terms of storage and retrieval of analyses:
  representation – search – counting – database
- In terms of data exchange with others:
  LAF-based connection with ongoing standardization work
Work on the disambiguation of -ung
Phenomena covered

- Analysis of small samples of nominalization data, for individual items (partly in cooperation with B4):
  - View: from corpus-observable data to sortal interpretation
  - Complementary to B4, which predicts sorts from verb semantics
  - Combination of both views seen as best overall strategy

- Modifiers (adjectives, relative clauses, participles, PPs) and Selectors (verbs or prepositions embedding -ung-words) as potential indicators:
  - Modifiers: laufende Messungen, Schätzungen im Mai
  - Selectors: Schätzungen zufolge, Messungen durchführen

- Support verb constructions (SVCs) as special contexts of -ung-nouns:
  - Identification of SVC candidates (cooperation with D2)
  - Identification of SVC candidate properties
Tools for disambiguating -ung
Prototype of a system: analysis – storage – search

• Analysis and storage:
  Use of the German grammar of the *translate* system
  (research version of Lingenio MT-product,
  cf. McCord 89, McCord 91, Eberle 02)
  – Internal representation uses underspecification,
    based on FUDRT (Eberle 2004)
  – Syntactic ambiguity represented by local underspecified structures

• Search:
  – Indicator function of context partners can be noted in the lexicon:
    Tool disambiguates -ung when enough knowledge is available
  – Large sets of sentences can be analyzed
  – In parallel: syntactic constellations of possible indicators
    can be extracted from the analyses
  – New aspects in terms of corpus tools:
    Deep analysis – underspecification –
    analysis and search in one tool
Tools for disambiguating -ung

Examples (Eberle/Heid 2008)

- Analysis, keeping track of indicator data:

  Messungen des Instituts in Gomel...

  $\{ l_1 : \text{institut}(i), l_2 : \text{in}(l_3 : \text{gomel}(g), L_{3'}_{x@ev|x@evorobj} ) \}$

  $l_{0x} : \begin{array}{c} \text{messung}^*(X) \\ \text{poss}(X,i) \end{array}$

  \{first(l1,l2) or first(l0,l2) \}

- Search for syntactic constellations: -ung fact

  $\{ l_{1u@eso} : \text{ung-deriv}(l_{1'e@event}' : L')_1, l_{2u'} : \text{mod}(L'_{2x@eso}), l_{3e''} : \text{fin}(e'') \}$

  \{⋄ first(l1,l2) \}

  $l_{0e} : L_0 \& \begin{array}{c} e \\ \text{subj}(e,u) \end{array}$

Results:


- 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.”
Storing and exchanging analyzed data
Relating B3 work with current corpus linguistic work

- Proposals for annotating syntactic ambiguity in LAF, the *Linguistic Annotation Framework* (Ide/Sudermann 2007)
  - dependency trees:
    addition of specific arcs to cater for e.g. ambiguous attachment
  - examples represented in XML (Kountz et al. 2008)
- Proposals for a database containing annotated sentences
  - explicit representation of syntactic ambiguity through local underspecification
  - fully searchable:
    both, fully specified and underspecified parts
  - data cover multiple layers of annotation:
    sentence – link to corpus – metadata – analyses
- Specification completed; implementation ongoing (Eberle, Eckart, Heid 2009)
Next steps in Phase I

- More work on -ung-data:
  Tools available, application to data extraction under way

- Tool improvements:
  - Broader coverage of translate-based analysis tool
  - Experiments with Schiehlen’s (2003) dependency parser
  - Work on reliability by means of a multi-engine approach:
    run two (several) analyzers in parallel
    → task: compare and combine analyses
    → method: use database formalism and functions

- Apply tools to other phenomena than just -ung, as suggested by reviewers: ambiguity in data extraction

- Database:
  - Implementation (postgresql)
  - Experiments to interface with Paula/Annis-infrastructure (SFB-632) and/or with inter-SFB project on sustainable corpora
Perspectives for Phase II
Plans to broaden and deepen B3’s programme of work

• Phenomena:
  Morphologically related words: verb – nominalization – SVC
    – Parallelisms and differences in subcategorization – selection – distribution
    – Prefix and particle verbs and their derived nominals

• Linguistic questions:
    – How can we use contexts of related items to better understand derivation processes?
    – How compositional are prefixing and particle verb formation?

• Corpus linguistic and tool questions:
    – How can we compare contexts for related items?
    – How can we expand multi-engine approach towards reliable underspecified analysis?