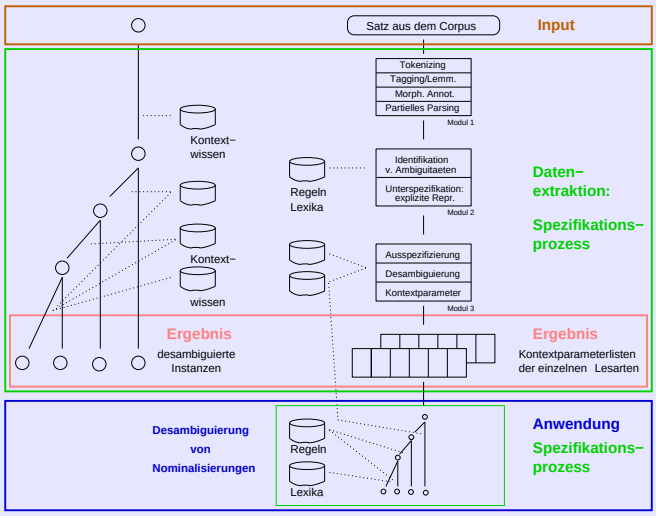
Unterspezifizierte Repräsentation und Disambiguierung sortal ambiger Nominalisierungen auf *-ung*

Kurt Eberle, Manuel Kountz, Ulrich Heid, Kerstin Eckart
Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung, Azenbergstraße 12, D-70174 Stuttgart

Aufgabenstellung

- Automatische Datenextraktion als *inkrementeller Spezifizierungsprozeß*:
  Berücksichtigung von Mehrdeutigkeiten
  bei der Extraktion von linguistischen Daten aus Textkorpora
  - **Ziele:**
    - Identifikation mehrdeutiger Ausdrücke eines vorgegebenen Typs in Texten
    - Identifikation von disambiguierungsrelevanter Information ('Indikatoren'):
      Welche Typen gibt es, welche Konstrukte finden sich im Kontext?
    - Verfahren zur Markierung von Ergebnissen der Korpussuche:
      Sind sie von Mehrdeutigkeiten "betroffen", d.h. nicht "sicher"?
    - Verfahren zur semi-automatischen Identifikation solcher Phänomene
  - **Ansatz:**
    - Unterspezifizierte Repräsentation von Mehrdeutigkeiten in Texten
    - Partielle Disambiguierung,
      wo dies für eine Extraktionsaufgabe nötig ist: Bootstrapping (s. "Methode")
- *Beispielfall:* sortale Ambiguität von *-ung*-Nominalisierungen:
  - *Absperrung*: Ereignis – Zustand – Objekt
    vgl. aber im FR.: *barrière*$_{\text{OBJEKT}}$ vs. *barrage*
  - Indikatoren aus dem Kontext:
    - z.B. Adjektive: *rot-weiße Absperrung*$_{\text{OBJEKT}}$
    - z.B. Verb vs. Subjekt: *die Absperrung*$_{\text{EREIGNIS}}$ *ist in 5 Minuten erfolgt*
    - z.B. Verb vs. Objekt: *die Absperrung*$_{\text{ZUSTAND}}$ *wird aufgehoben*
- Halbautomatische *Extraktion neuer disambiguierungsrelevanter Informationen*:
  Erweiterung des Indikatoren-Inventars um zusätzliche lexikalische und strukturelle Eigenschaften



Methode: Deterministisches Parsing mit "Ambiguity Awareness"

- Eindeutige **unterspezifizierte Satzanalyse** (Syntax und Semantik):
  vermeidet "forced guessing" (vgl. KOUNTZ ET AL. 2007)
  Ergebnis:
  - Sichere, eindeutige syntaktische Dependenzanalyse mit integrierter Ambiguitätsbeschreibung
  - Eindeutige unterspezifizierte semantische Repräsentation:
    Flache unterspezifizierte Diskursrepräsentationsstrukturen (FUDRS, EBERLE 2004)
- Aufgabenbezogene deterministische **partielle Disambiguierung**
  mittels Kontextwissen, z.B. mittels Selektionsrestriktionen (vgl. SPRANGER/HEID 2007)
  Ergebnis:
  - Menge aufgabenbezogen disambiguierter unterspezifizierter Repräsentationen
- **Bootstrapping** zur Erweiterung des Indikatoren-Inventars
  1. Korpussuche: unterspezifizierte Suchanfragen
  2. Extraktion von Indikatoren-Kandidaten: partielle Disambiguierung, wo nötig
  3. Manuelle Sichtung,
     Erfassung von Indikatoren und Analysen der *-ung*-Nominalisierungen in Lexikon und Grammatik
  4. Verbesserte Extraktion

Aktuelle Implementierung

Erweiterung der Analysekomponente des kommerziellen maschinellen Übersetzungssystems *translate* (LINGENIO (`http://www.lingenio.de`), MCCORD 1989)

- **Grammatik:**
  Umfangreiche Dependenzgrammatik (Deutsch, auch: Französisch, Englisch):
  unifikationsbasierter Formalismus der Slot Grammar (MCCORD 1991)
- **Unterspezifizierte Repräsentation:**
  - *Syntax:* Dependenzstrukturen
  - *Semantik:* FUDRSen: Flache unterspezifizierte Diskursrepräsentationsstrukturen,
    (EBERLE 2001, EBERLE 2004)
- **Ausgabe:**
  Für die jeweilige Extraktionsaufgabe relevante Lesarten:
  z.B. Nominalisierung plus sortale Ausdifferenzierung: *Absperrung*$_{\text{EREIGNIS}}$ vs. *Absperrung*$_{\text{OBJEKT}}$,
  sowie effizient durchsuchbare Sammlung von Indikatoren aus Texten

Analyse- und Ausgabebeispiele

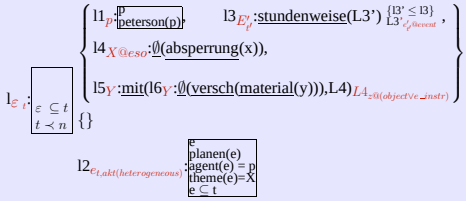
Peterson plante stundenweise Absperrungen mit verschiedensten Materialien.

(I) STUNDENWEISES PLANEN VON ABSPERRUNGEN, DIE AUS VERSCHIEDENSTEN MATERIALIEN BESTEHEN
(Absperrung als Objekt: *o*)

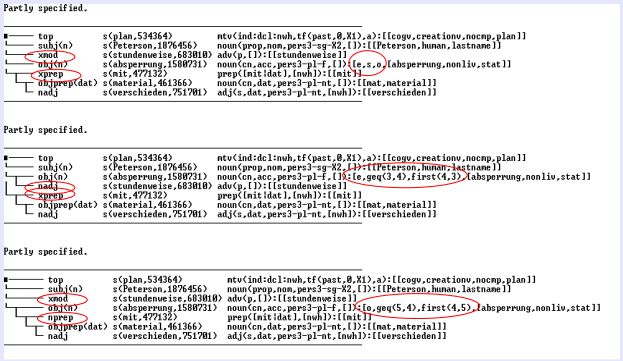
(II) PLANEN VON STUNDENWEISEN ABSPERRUNGEN
(Absperrung als Ereignis: *e*)

(III) STUNDENWEISES, MIT VERSCHIEDENSTEN MATERIALIEN DURCHGEFÜHRTES PLANEN VON ABSPERRUNGEN
(ohne Disambiguierung: *eso*)

Flache unterspezifizierte DRS (FUDRS)
mit Kern-DRS und Menge von Modifikatoren und (relevanten) DRF-'Dekorationen'



Disambiguierungsresultate der erweiterten *translate*-Analyse:



Nächste Ziele

- **Meta-Analysewerkzeug:**
  - Einbeziehung und Anpassung verschiedener Analysewerkzeuge
  - Architektur für die Integration der Werkzeuge: Möglichkeit zur Kombination von Einzelresultaten,
    Ziel: Optimierung der Verläßlichkeit der Analysen
- **Datenbank** zur Verwaltung von Sätzen und ihren Analysen:
  Forschungswerkzeug: Suche in Analysen, z.B. für Bootstrapping
- **Abbildung** der unterspezifizierten Repräsentation
  auf *Korpusrepräsentations-Standards* (Auf Basis LAF/GrAF, z.B. KOUNTZ ET AL. 2008)
- **Ausweitung** des Ansatzes auf Daten aus anderen Sprachen (EN, FR),
  ggf. Nutzung von Übersetzungswissen (Paralleltext) zur Disambiguierung:
  Die Bedienung ist schwierig vs. *la serveuse / l'utilisation est difficile*

Projekthintergrund

- Sonderforschungsbereich 732
  - Thema: *Incremental Specification in Context*
  - Laufzeit: 1. Juli 2006 – 31. Juni 2010
- Projekt B3: *Disambiguierung von Nominalisierungen bei der Extraktion linguistischer Daten aus Corpora*

URLs

- SFB 732: `http://www.uni-stuttgart.de/linguistik/sfb732/`
- Lingenio: `http://www.lingenio.de`

Literatur

[Eberle2001] Kurt Eberle. 2001. FUDR-based MT, head switching and the lexicon. In: *Proceedings of the eighth Machine Translation Summit, Santiago de Compostela.*

[Eberle2004] Kurt Eberle. 2004. Flat underspecified representation and its meaning for a fragment of German. *Habilitationsschrift*, Stuttgart.

[McCord1989] Michael McCord. 1989. Design of LMT. In: *Computational Linguistics*, 15(1), pp33–52.

[McCord1991] Michael McCord. 1991. The Slot Grammar System. In: Wedekind, J. & Rohrer, Ch. (eds.): *Unification in Grammar*, MIT-Press.

[Kountz et al.2007] Manuel Kountz, Ulrich Heid, Kristina Spranger. 2007 Automatic sortal Interpretation of German Nominalisations with *-ung*: Towards using underspecified Representations in Corpora. In: Proceedings of *Corpus Linguistics 2007*

[Kountz et al.2008] Manuel Kountz, Ulrich Heid, Kerstin Eckart. 2008 A LAF/GrAF based Encoding Scheme for underspecified Representations of Dependency Dtructures. erscheint in: Proceedings of the *Language Resources and Evaluation Conference (LREC)*, Marrakesch

[Spranger/Heid2007] Kristina Spranger, Ulrich Heid. 2007 Applying Constraints derived from the Context in the Process of Incremental Sortal Specification. In: *Proceedings of the Workshop CSLP@Context 2007*, Roskilde.