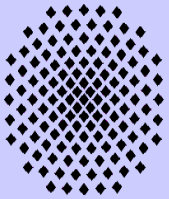


Eine Datenbank als multi-engine

für Sammlung, Vergleich und Berechnung möglichst verlässlicher unterspezifizierter syntaktisch/semantischer Satzrepräsentationen

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

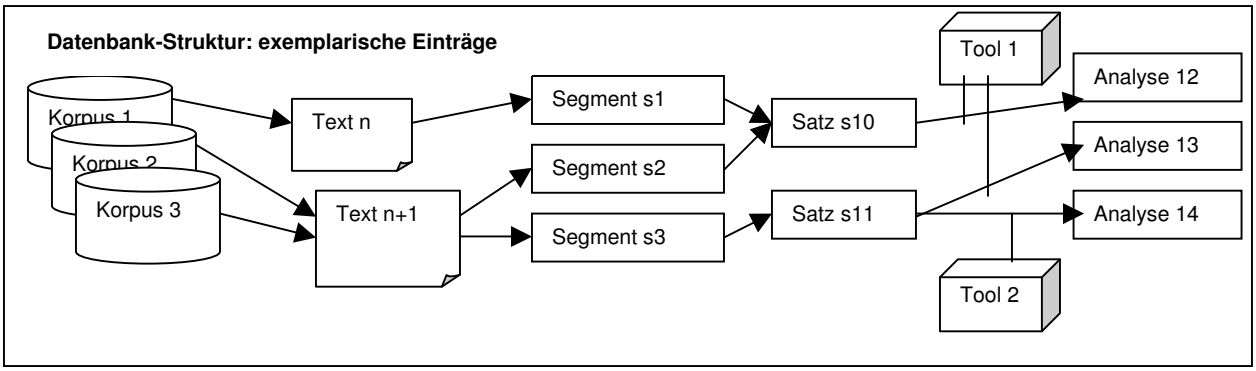
Handout

Ziele

- Möglichst verlässliche Analysen von Sätzen aus großen Korpora
  als Basis für die Überprüfung linguistischer Hypothesen zu verschiedenen Fragestellungen aus Syntax und Semantik
- Verwaltung von Ergebnissen unterschiedlicher Analyseverfahren
- Vergleich von Analysen
  - zur Berechnung von Verlässlichkeitswerten
  - als Infrastruktur zur Optimierung von Analysewerkzeugen
- Berechnung von Analyse-Integrationen
- Kompatibilität mit der ANNIS-Datenbank-Konzeption und dem PAULA-Repräsentationsformat (SFB632/D1, vgl. Chiarcos et al)

Anforderungen an die Datenbank

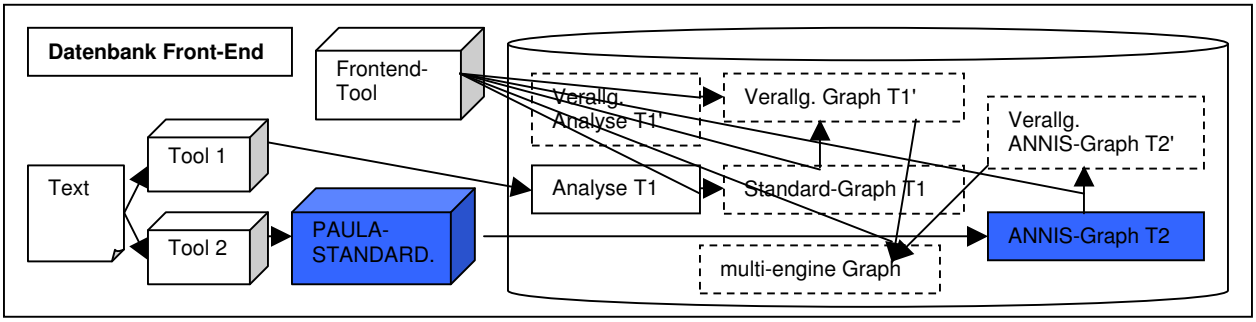
- Speicherung von Korpusdaten und -analysen verschiedenster Art
- Verfügbarkeit unterschiedlicher Repräsentationen der Analysen
  - für möglichst einfache Inspektion der Analysen
  - für möglichst notationsunabhängigen Vergleich von unterschiedlichen Analysen und - indirekt - Analysewerkzeugen
- Verfügbarkeit von Relationen zwischen Repräsentationen von Analysen (für Subsumtion, Ähnlichkeit, etc.)
- Reproduzierbarkeit von Analysen durch datenbank-interne Verwaltung der Analysewerkzeugdaten
- Konzeption als **temporale** Datenbank zur Administration von Dateninterpretationen: `created, invalidated`
  - zur Abschätzung des Qualitätsverlaufs sich verändernder Werkzeuge
  - für detaillierte Verlässlichkeitsaussagen (Hypothese z.B.: neue Zusammenstellungen verlässlicher als ältere)
- Intelligente Datenbank-Schnittstelle zur Berechnung von
  Verallgemeinerungen, Vergleichen und Integrationen von Analysen
- Annotation von Satzanalysen mit Vergleichsinformation und daraus abgeleiteter Verlässlichkeitsinformation



Datenbank-Konzeption

- Unterscheidung zwischen Makro- und Mikroebene:
  - Analysen sind atomare Objekte der Makroebene
  - aber strukturierte Objekte (Graphen) mit Knoten und Kanten auf der Mikroebene
- Unterscheidung zwischen aktuellen und vergangenen Analysezuständen
- Reproduzierbarkeit vergangener Zustände
- Makro- und Mikroebene bestehen jeweils im Wesentlichen aus Objekten und zweistelligen gerichteten Relationen über Objekten
- Objekte und Relationen sind nach atomaren Typen und Boole'schen Kombinationen von Typen klassifiziert; sie können zusätzlich über Attribut-Wert-Paare beschrieben sein (wie in einer Feature-Logik mit Subtypen)

Relationen zwischen Analysen und Analyse-Graphen



Nächste Ziele

- Berücksichtigung einer Reihe weiterer Analyse-Tools (neben BitPar, FSPar (vgl. Schiehlen 2003) u.a.)
- Integration von TIGER-Daten
- Verfeinerung der Annotationsterminologie (vgl. PAULA-Ontologie; Dipper et al 2007)
- Erweiterung der Verallgemeinerungs- und Überdeckungsverfahren

Beispielanalysen

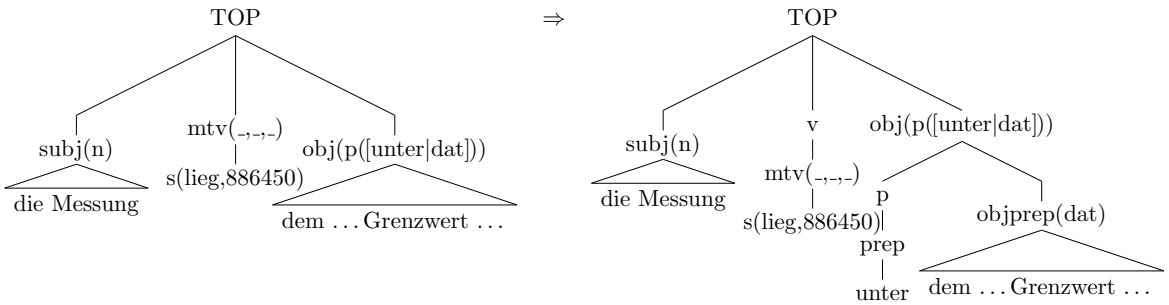
- Beispielsatz: *Auch bei den CO-Werten liegen die Messungen weit unter dem zulässigen Grenzwert von 250 ppm (parts per million, Bestandteile in einer Million Teile), von dem an ein Tunnel gesperrt werden muss; die bei Dauerbetrieb allgemein zulässigen Werte liegen zwischen 100 bis 150 ppm.*

- Dependenz-Analyse (Lingenio-Tool)

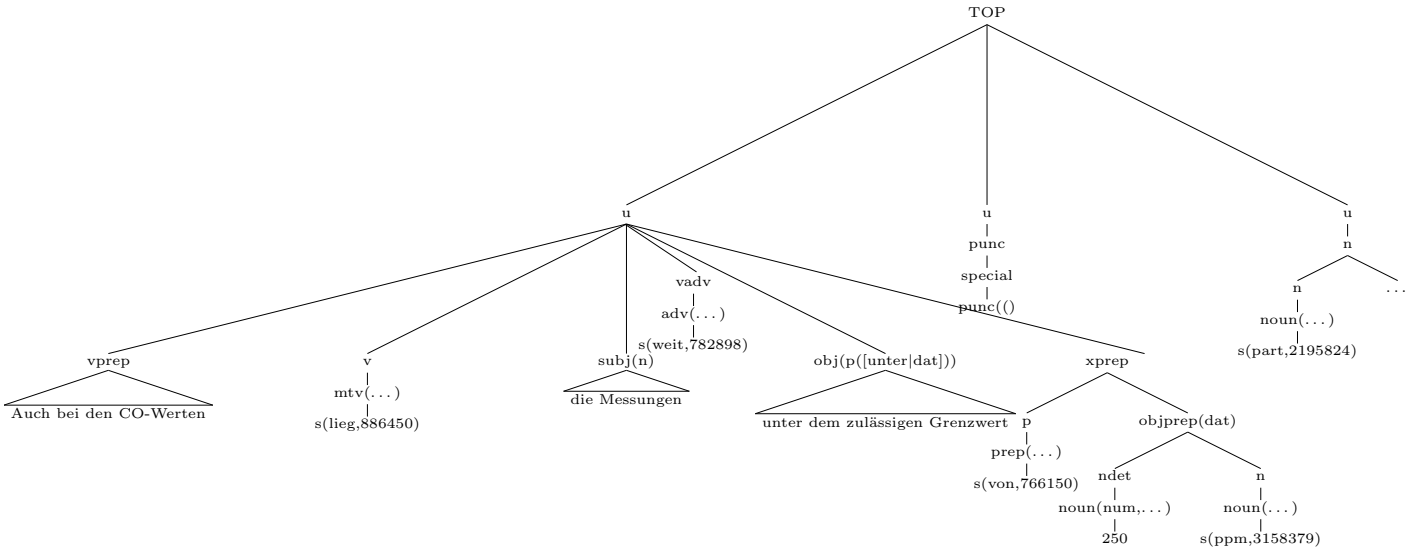
```
o------- top                  0                    incomplete:[[]]
‘------- u                    s(lieg,886450)       mtv(dep:dcl:nwh,tf(pres,0,X1),a):[[lieg,locv]]
| ‘----- vprep                s(bei,81024)         prep([bei|dat],[nwh]):[[bei]]
| | ‘--- padv                 s(auch,50971)        adv(p,[]):[[auch,discrel_adv,dsa,focus_adv,ppadv]]
| | ‘--- objprep(dat)         s(wert,786921)       noun(cn,dat,pers3-pl-m,[]):[[mass,quant,wert]]
| |   ‘- ndet                 s(den,d)             det(dat,pers3-pl-m,[def]):[[d,den]]
| |   ‘- ncompound            s(CO,2332416)        noun(cn,cmp,pers3-sg-nt,[]):[[CO,gaso]]
| ‘----- subj(n)              s(messung,470364)    noun(cn,nom,pers3-pl-f,[]):[[activity,messung]]
| | ‘--- ndet                 s(die,d)             det(nom,pers3-pl-f,[def]):[[d,die]]
| ‘----- vadv                 s(weit,782898)       adv(p,[focus(14),adv]):[[lmj,weit]]
| ‘----- obj(p([unter|dat])) s(grenzwert,304915)  noun(cn,acc,pers3-sg-m,[]):[[grenze0,grenzwert,quant]]
| | ‘--- ndet                 s(dem,d)             det(acc,pers3-sg-m,[def]):[[d,dem]]
| | ‘--- nadj                 s(zulässig,817163)   adj(p,acc,pers3-sg-m,[]):[[ncomp,nsup,zulässig]]
| ‘----- xprep                s(von,766150)        prep([von|dat],[nwh]):[[von]]
|   ‘--- objprep(dat)         s(ppm,3158379)       noun(cn,dat,pers3-pl-nt,[]):[[measure,ppm]]
|     ‘- ndet                 250                  noun(num,dat,pers3-pl-nt,[num]):[[250]]
‘------- u                    punc(()              special:[[]]
‘------- u                    s(part,2195824)      noun(cn,[nom,dat,acc,gen],pers3-pl-m,[]):[[part]]
| ‘----- nprep                s(per,526321)        prep([per|acc],[nwh]):[[per]]
| | ‘--- objprep(acc)         coord(com,s(million,474660),s(bestandteil,93963)) noun(cn,acc,pers3-pl-X2,[]):[[]]
| |   ‘- lconj                s(million,474660)    noun(cn,acc,pers3-sg-f,[]):[[million,num,perco(nobj)]]
| |   ‘- rconj                s(bestandteil,93963) noun(cn,acc,pers3-pl-m,[]):[[bestandteil,oreal,part]]
| ‘----- xprep                s(in,353022)         prep([in|dat],[nwh]):[[in,loc_prep]]
|   ‘--- objprep(dat)         s(million,474660)    noun(cn,dat,pers3-sg-f,[]):[[million,num,perco(nobj)]]
|     ‘- ndet                 s(ein,182125)        det(dat,pers3-sg-f,[indef]):[[ein]]
‘------- u                    s(teil,694897)       noun(cn,[nom,acc],pers3-pl-m,[]):[[nonliv,nonself,part,teil,perco(nobj
‘------- u                    punc())              special:[[]]
‘------- u                    punc(,)              special:[[]]
‘------- u                    s(müss,488021)       mtv(dep:dcl:nwh,tf(pres,0,X3),a):[[einstellung0,modv,müss]]
  ‘----- subj(n)              s(tunnel,715317)     noun(cn,nom,pers3-sg-m,[]):[[tunnel]]
  | ‘--- ndet                 s(ein,182125)        det(nom,pers3-sg-m,[indef]):[[ein]]
  ‘----- auxcomp(binf)        s(sperr,652376)      mtv(dep:inf:nwh,tf(inf,0,X4),p):[[sperr]]
    ‘--- subj(n)              s(dem,d)             noun(pron(defprn),dat,pers3-sg-X5,[]):[[d,dem]]
    | ‘- npbrk                s(an,26529)          adv(p,[]):[[an,nopre]]
    ‘--- obj(n)               empty                coref(34)
```

  - u – 'unknown': Beitrag der so bezeichneten Struktur zum Gesamtsatz kann nicht ermittelt werden

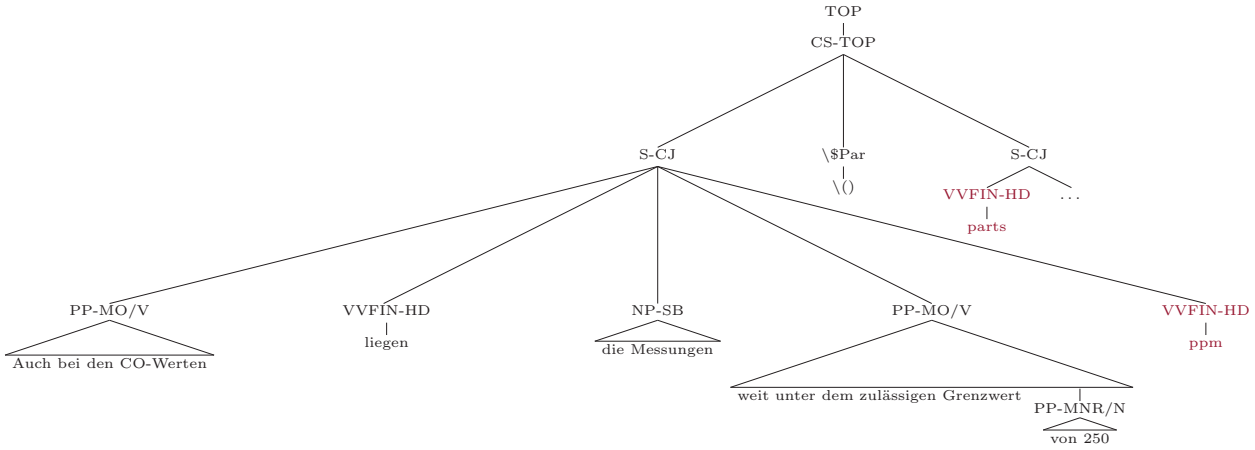
- Abbildung: Dependenz-Analyse (Lingenio-Tool) → Konstituentenstruktur-Graph
  - Erweiterung durch Einfügen von Projektionsknoten für Kopf-Elemente (vgl. Eberle 2002)
  - Beispiel:



- Konstituentenstruktur-Graph der Dependenz-Analyse (Ordnungsstruktur der zugeordneten FUDRS)

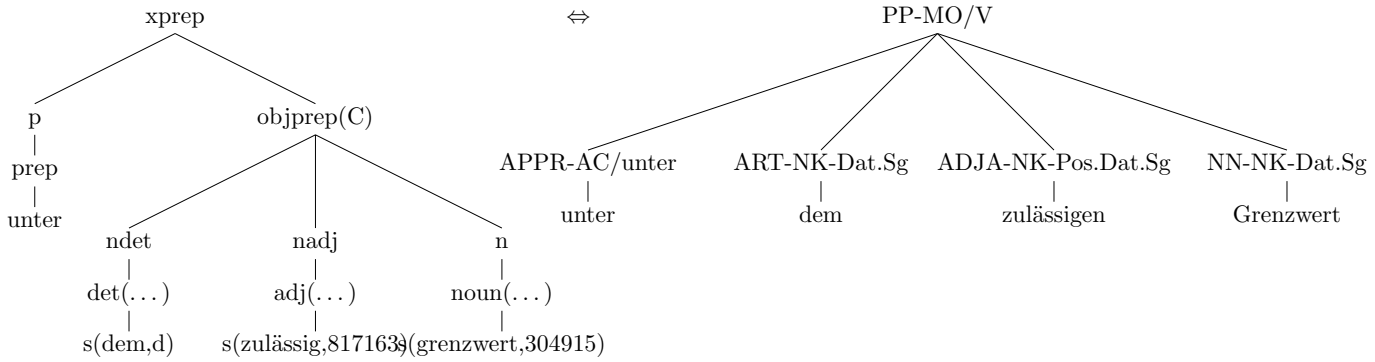


- Alternative Konstituentenstruktur-Analyse (BitPar-Tool, vgl. Schmid 2004)

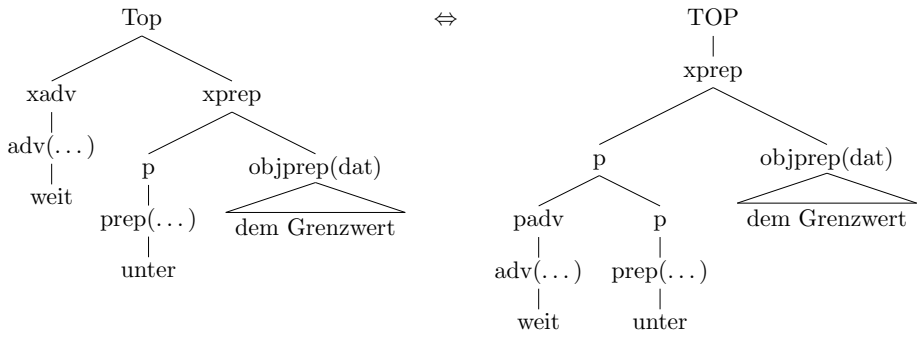


- Vergleich: Ähnlichkeiten und Unterschiede zwischen Lingenio und BitPar
  - Analysen sind bis zur öffnenden Klammer relativ ähnlich: *Auch bei ... von 250 ppm (...*
    Unterschied bei der Interpretation von *ppm:* Lingenio-Tool kennt das Wort, BitPar nicht
    → Lingenio-Analyse verlässlicher
  - Unterschied bei der Interpretation von *weit:*
    Als Adverbiale in Lingenio-Analyse, Präp.-Modifikation in BitPar
    → häufiger, relativ unbedeutender Unterschied
  - Unterschied bei der Interpretation der *von-PP:*
    Als unterspezifiziert verankerte PP in Lingenio-Analyse, als Modifikation von *Grenzwert* in BitPar
    → häufiger struktureller Unterschied: ändert nichts an kategorialer oder Teilstruktur-Interpretation
  - Unterschied bei der Interpretation von *CO-Wert:*
    Analytisch als Kompositum in Lingenio-Analyse, als atomarer Ausdruck in BitPar
    → häufiger, relativ unbedeutender Unterschied
  - Regelhafter Unterschied der strukturellen Repräsentation von PPs:
    flach in BitPar-Analyse, z.B.: **p det adj n**, strukturiert in Lingenio-Analyse: **p objprep (det adj n)**

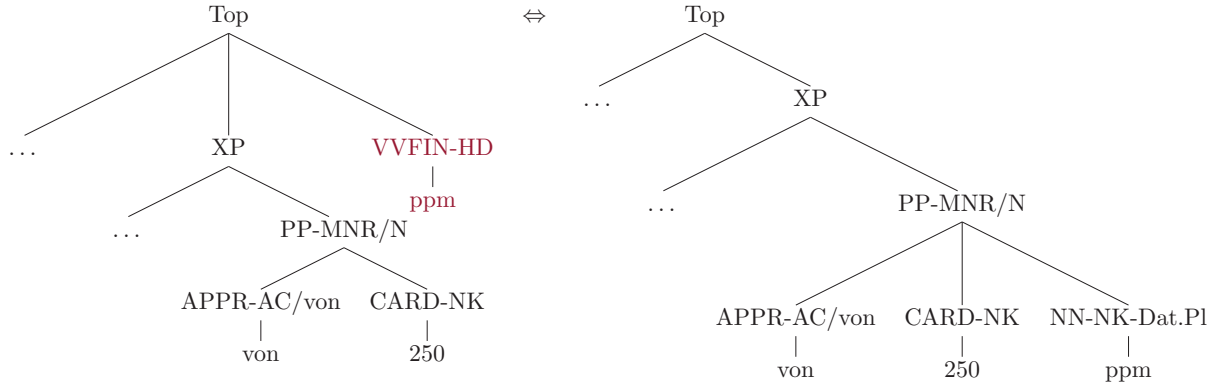
- Graph-Äquivalenzregeln
  Beispiel: PP-Repräsentationskonventionen



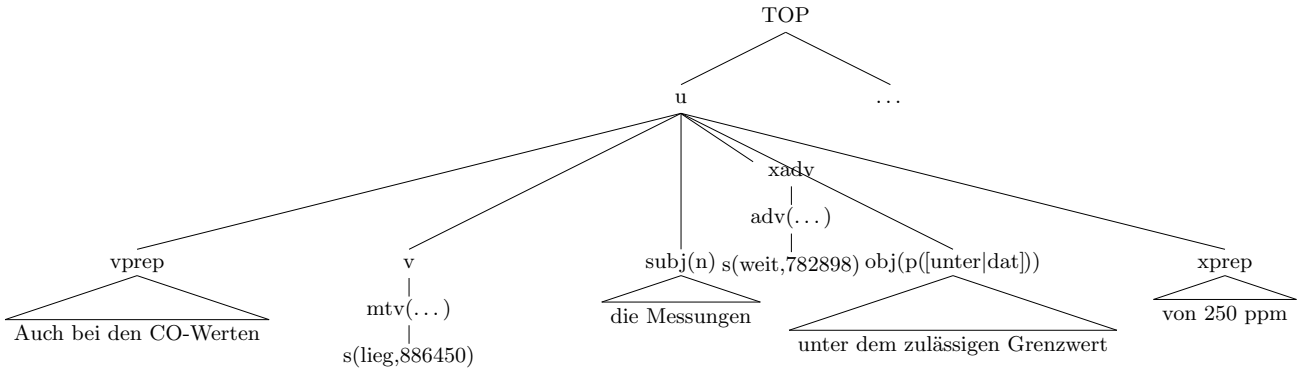
- Minimale Interpretationsdifferenz-Regeln
  Beispiel: Adverb als Adverbial- versus als Skalar-Modifikator



- Begründete Reparatur-Regeln
  Beispiel: Korrigiere uninformierte Interpretation entsprechend der informierten Repräsentation



- Gemeinsamer Teilgraph modulo Äquivalenz und minimaler (d.h. nur lokaler) Interpretationsdifferenz



- Verlässlich identifizierte potentielle Selektoren und Modifikatoren der -ung-Nominalisierung: V-liegen, unter-PP, von-PP

Rahmen

- Sonderforschungsbereich 732: Incremental Specification in Context
- Projekt B3: Disambiguierung von Nominalisierungen bei der Extraktion linguistischer Daten aus Corpustext
- Laufzeit: 1. Juli 2006 – 30. Juni 2010

URLs

- Lingenio: `http://www.lingenio.de`
- SFB 732: `http://www.uni-stuttgart.de/linguistik/sfb732/`
- TIGER: `http://www.coli.uni-saarland.de/projects/tiger/`

Literatur

[**Chiarcos et al 2008**] C. Chiarcos, S. Dipper, M. Götze, J. Ritz, M. Stede. 2008. A Flexible Framework for Integrating Annotations from Different Tools and Tagsets. In: Proceeding of the Conference on Global Interoperability for Language Resources, Hong Kong, January 2008

[**Dipper et al 2004**] S. Dipper, M. Götze, M. Stede, T. Wegst. 2004. ANNIS: A Linguistic Database for Exploring Information Structure. In: S. Ishihara, M. Schmitz, A. Schwarz (eds.): Working Papers of the SFB632, Interdisciplinary Studies on Information Structure, ISIS, 1 (245-279). Universitätsverlag Potsdam, Potsdam

[**Dipper 2005**] S. Dipper. 2005. XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In: Proceedings of Berliner XML Tage 2005, BXML 2005, pp. 39-50, Berlin

[**Dipper et al 2007**] S. Dipper, M. Götze, S. Skopeteas (eds.). 2007. Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax,Semantics, and Information Structure. Volume 7 of Interdisciplinary Studies on Information Structure, ISIS. Universitätsverlag Potsdam, Potsdam

[**Eberle 2002**] K. Eberle. 2002. Tense and Aspect Information in a FUDR-based German French Machine Translation System. In: H.Kamp, U.Reyle (eds.) How we say WHEN it happens. Contributions to the theory of temporal reference in natural language. Ling. Arbeiten, Band 455, Niemeyer, Tübingen

[**Eberle/Eckart 2009**] K. Eberle, K. Eckart. 2009. Eine Datenbank für Textanalysen – Design und Beispiele. Institut für Maschinelle Sprachverarbeitung. Universität Stuttgart. Manuskript

[**Eberle et al 2008**] K. Eberle, U. Heid, M. Kountz , K. Eckart. 2008. A Tool for Corpus Analysis using partial Disambiguation and Bootstrapping of the Lexicon. In: A. Storrer et al. (eds.): Text Resources and Lexical Knowledge: Selected Papers from the 9th Conference on Natural Language Processing KONVENS 2008

[**Roßdeutscher 2007**] A. Roßdeutscher. 2007. Syntactic and Semantic constraints in the formation and interpretation of -ung-nouns. Vortrag beim Workshop: Nominalizations across Languages, Stuttgart, 11./12. Dezember 2007

[**Schiehlen 2003**] M. Schiehlen. 2003. A Cascaded Finite-State Parser for German. In: Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics, EACL'03, pp. 163-166 Budapest, Hungary

[**Schmid 2004**] H. Schmid. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In: Proceedings of the 20th International Conference on Computational Linguistics, Coling'04, volume 1 pp. 162-168 Geneva, Switzerland

[**Spranger/Heid 2007**] K. Spranger, U. Heid. 2007. Applying Constraints derived from the Context in the process of Incremental Sortal Specification of German -ung-Nominalizations. In: Proceedings of the 4th Int. Wkshp. on Constraints & Lang. Processing, CSLP