Auf dem Weg zu Normen für die Repräsentation annotierter Korpora

Gottfried Herzog und Normenausschuss NA-105-00-06 AA
DIN – Deutsches Institut für Normung, Burggrafenstraße 6, 10787 Berlin

Wie erfolgt Normungsarbeit?

- Wieso Normen für Sprachressourcen?
  - Best Practice: vermeidet Doppelarbeit
  - Nachhaltigkeit:
    Pflege – Wartung – multiple Nutzbarkeit
  - Austauschformat:
    Kombinierbarkeit – Austausch von Ressourcen
- Wie entstehen Normungsvorschläge?
  - Sichtung vorhandener Formate und Tools
  - Vorschläge – Änderungen – Updates – ...
  - Stufenweise Verbesserung
- Wer ist involviert?
  - International: International Organisation for Standardisation, ISO:
    Strukturen für internationale Abstimmung,
    Organisation der Schritte bis zur endgültigen Norm
    Relevante Arbeitsgruppe: ISO TC37/SC-4
  - National: DIN-Normenausschuss NA-105 ("Terminologie"):
    - Bindeglied zu ISO (Gremienbetreuer: Gottfried Herzog, DIN, Berlin)
    - Organisation, Administration, Berichtswesen
  - Arbeitsausschuss *Sprachressourcen* im NA-105: NA-105-00-06 AA
    Spiegelgruppe zu ISO TC37/SC-4:
    Formuliert deutsche Stellungnahmen bzw. Vorschläge zu Normen
  - Mitarbeitende: abgeordnet von Firmen und Universitäten
    - Teilnahme ist freiwillig, auch möglich für einzelne Themenbereiche
    - Mitwirkung durch Interessenbekundung, Ernennung, aktive Mitarbeit
    - Kreis ist offen: es zählt Expertise

Normen für Sprachressourcen

- Aspekte – Abstraktionsebenen:
  - Theoretische Sicht auf Prinzipien der Modellierung
  - Methoden der Modellierung

  } Metamodell
  - Spezifische Modellierungsinstanzen
    - für einzelne linguistische Beschreibungsebenen
    - Datenkategorien (Definition und Dokumentation)
  - Beispielanwendungen
- Überblick

	Repräsentation	Phänomene in Corpora	Lexika	Terminologie
Metamodelle – Modellierungsprinzipien – Modellierungsmethoden	FSD: Feature System Declarations	LAF/GrAF	LMF: Lexical Markup Framework	TMF: Terminology Markup Framework
Spezif. Modellierungen – Datenkategorien – Einzelebenen	DCR FSR: Feature Structure Representations	DCR MAF, SynAF, SemAF	DCR	DCR: Data Category Registry TBX: Term. Exchange Format
Anwendungsbeispiele – Einzelebenen		(Experimente)		IATE-Wörterbuch

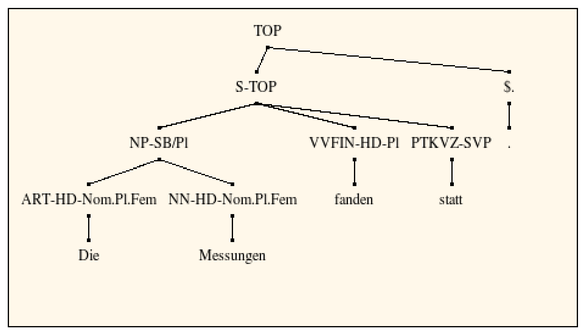
- Modellierungsprinzipien
  - LAF: Linguistic Annotation Framework – GrAF: Graph Annotation Format (ISO/DIS 24612)
    - Repräsentation von linguistischen Daten und Annotierungen
- Modellierungsmethoden
  - Repräsentation durch Merkmalsstrukturen: FSD: Feature System Declarations (ISO/DIS 24610-2)
  - Lexika: LMF: Lexical Markup Framework (ISO 24613:2008)
- Spezifische Modellierungen
  - Verfahren zur Definition von Datenkategorien: DCR (ISO 12620: 2009)
  - Einzelne Ebenen der linguistischen Beschreibung:
    - MAF: Morphosyntactic Annotation Framework (ISO/DIS 24611)
    - SynAF: Syntactic Annotation Framework (ISO/DIS 24615)
    - SemAF: Semantic Annotation Framework (ISO/DIS 24617):
      einzelne Phänomene, z.B. Temporalausdrücke
  - Repräsentation: FSR: Feature Structure Representation (ISO 24610-1:2006)

Normen für Korpora: Grundideen

- Graphbasierte Repräsentation von Text und Annotationen: stand-off XML
- Mehrschichtiges Modell: erweiterbar
- Gemeinsame Datenkategorien: ISOcat als Inventar und Beschreibungstool
- Repräsentation von analysierten Korpusdaten, z.B. annotierten Bäumen:
  - Wortformen: MAF
  - Syntaktische Struktur: SynAF
  - Semantische Annotationen: SemAF
- Metamodell (als allgemeines Austauschformat): LAF/ GrAF

Beispiel für eine Korpusannotation: Kodierung in LAF

- Ausgangspunkt: Analyse von BitPar (Schmid 2004): *Die Messungen fanden statt.*



- BitPar Input als *BaseSegmentation*
  Referenzierung der textuellen Primärdaten: Character-Offsets

```
|D|i|e| |M|e|s|s|u|n|g|e|n| |f|a|n|d|e|n| |s|t|a|t|t|.|
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7
                  1                   2
```

LAF/GrAF-Element `region` – *SynAF-Subset : token*

```
<graf:region id=''r1'' anchors=''0 3''/>   Die
<graf:region id=''r2'' anchors=''4 13''/>  Messungen
<graf:region id=''r3'' anchors=''14 20''/> fanden
<graf:region id=''r4'' anchors=''21 26''/> statt
<graf:region id=''r5'' anchors=''26 27''/> .
```

- Annotation: Terminalknoten

```
<graf:node id =''n1''>
  <graf:link to=''r1''/>
  <graf:as type=''BitPar''>
    <graf:a label=''msd''>
      <ns1:fs>
        <ns1:f name=''pos''>
          <symbol value=''ART''/>
        </ns1:f>
        <ns1:f name=''case''>
          <symbol value=''Nom''/>
        </ns1:f>
        <ns1:f name=''number''>
          <symbol value=''Pl''/>
        </ns1:f>
        ...
```

TerminalKnoten
Referenz auf Region r1: ‘‘Die’’
Annotation Set
morpho-syntactic description
Feature Structure
Features

(ART-HD-Nom.Pl.Fem Die)

Annotation: Nicht-Terminalknoten

```
<graf:node id =''n6''>
  <graf:as type=''BitPar''>
    <graf:a label=''cd''>
      <ns1:fs>
        <ns1:f name=''cat''>
          <symbol value=''NP''/>
        </ns1:f>
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:node>
```

Nicht-TerminalKnoten
(kein Link)
constituency description

NP-Knoten

```
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:node>
```

- Annotation: Kanten

```
<graf:edge id=''e4'' from=''n7'' to=''n3''>
  <graf:as type=''BitPar''>
    <graf:a label=''cd''>
      <ns1:fs>
        <ns1:f name=''role''>
          <symbol value=''HD''/>
        </ns1:f>
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:edge>
```

Kante von Knoten n7 (S) nach n3 (VVFIN:‘‘findet’’)

```
<graf:edge id=''e5'' from=''n7'' to=''n4''>
  <graf:as type=''BitPar''>
    <graf:a label=''cd''>
      <ns1:fs>
        <ns1:f name=''role''>
          <symbol value=''SVP''/>
        </ns1:f>
      </ns1:fs>
    </graf:a>
  </graf:as>
</graf:edge>
```

Kante von Knoten n7 (S)nach n4 (PTKVZ:‘‘statt’’)

Kanten werden auf die selbe Art annotiert wie Knoten
Alternative: Zusätzliche Annotationsknoten einführen

- Relevante Aspekte aus Sicht der Computerlinguistik:
  - LAF ist nur als Austauschformat gedacht (Meta-Format)
  - Beliebige tool-spezifische Repräsentationen können in LAF rekodiert werden
  - Mittelfristig sollen MAF und SynAF für Korpuskodierung benutzt werden:
    noch in der Erprobung
  - Ressourcenprojekte (z.B. CLARIN, D-SPIN, FlareNet,...) streben langfristig an:
    Interfaces zu MAF/SynAF/SemAF...

Information

ISO 12620:2009 Terminology and other language and content resources –
Specification of data categories and management of a Data Category Registry for language resources. 2009.
ISO 16642:2003 Computer applications in terminology – Terminological markup framework. 2003.
ISO 24610-1:2006 Language rsource management - Feature structures - Part 1: Feature structure representation. 2006.
ISO/DIS 24610-2 Language resource management – Feature structures – Part 2: Feature system declaration. 2009.
ISO/DIS 24611 Language resource management - Morpho-syntactic annotation framework (MAF). 2008.
ISO/DIS 24612 Language resource management - Linguistic annotation framework (LAF). 2009.
ISO 24613:2008 Language resource management - Lexical markup framework (LMF). 2008.
ISO/DIS 24615 Language resource management - Syntactic annotation framework (SynAF). 2009.
ISO/DIS 24617 Language resource management – Semantic annotation framework (SemAF). 2009.
ISO 30042:2008 Systems to manage terminology, knowledge and content – TermBase eXchange (TBX). 2008.

Schmid (2004) H. Schmid "Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors", in: *Proceedings of the 20th International Conference on Computational Linguistics* (COLING 2004), Geneva, Switzerland.

Interoperabilität: A. Witt, U. Heid, F. Sasaki, G. Sérasset: "Multilingual Language Resources and Interoperability", in: *Language Resources and Evaluation* (2009) 43: 1 – 14 [Einleitung zum Themenheft der Zeitschrift].
Prinzipien, Featurestrukturen: T. Trippel, T. Declerck, U. Heid: "Sprachressourcen in der Standardisierung", in: *LDV-Forum* 20:2 (2005): 17 – 29 [Themenheft Korpuslinguistik].

DIN: http://www.din.de/
DIN-NA 105: http://www.nat.din.de/
ISO: http://www.iso.org/
ISOcat: http://www.isocat.org/