Standards in der linguistischen Annotation

– Aktuelle Normen und De-facto-Standards –

Maik Stührenberg

13.06.2012

Gliederung

Standards und Standards

Standards im originären Sinne lassen sich in zwei Kategorien unterscheiden:

De-jure-Standards oder auch Normen

Die Spezifikation wird durch ein Standardisierungsgremium erarbeitet, das einen gesetzlichen Auftrag hat (z. B. Deutsche Institut für Normung, DIN e. V., oder die International Organization for Standardization, ISO)

De-facto- oder auch Quasi-Standards

Die Spezifikation wird durch eine Firma oder eine andere Organisation ohne gesetzlichen Auftrag erarbeitet; die Reputation als Quasi-Standard erfolgt durch eine entsprechende Nutzung (z. B. durch Marktmacht oder aber durch Übereinstimmung der Nutzer)

Standards in der linguistischen Annotation

Im Bereich der linguistischen Annotation ist zu unterscheiden zwischen grundlegenden und darauf aufbauenden Spezifikationen

Beispiele für grundlegende Standards

- Unicode als Standard für die Zeichenkodierung
- Extensible Markup Language (XML) als Metasprache zur Definition von Auszeichnungssprachen
- Grammatikformalismen zur Definition eigener Dokumentgrammatiken, wie XML DTD, XML Schema Description (XSD) oder RELAX NG

Beispiele für darauf aufbauende Standards

- Konkrete Auszeichnungssprachen zur Annotation linguistischer Phänomene, z. B. die TEI Guidelines oder das Linguistic Annotation Framework (LAF)

Das Konzept der Annotation

Annotation:

- Anreicherung von Informationen an einen Untersuchungsgegenstand
- Sowohl die zu annotierenden Daten (Primärdaten) als auch die Komponenten des Annotationsinventars liegen in textueller Repräsentation vor
- Eine spezielle Syntax trennt Annotation und Primärdaten
- Syntax und Annotationsinventar werden durch eine Auszeichnungssprache definiert

Das Konzept der Annotation

Beispielkonstruktion

Hey Paul! Would you give me
the hammer?

Beispiel für eine POS-Annotation (Stanford-NLP-Tagger, txt-Ausgabe)

Hey_NNP Paul_NNP !_.
Would_MD you_PRP give_VB me_PRP the_DT hammer_NN ?_.

Das Konzept der Annotation

Beispiel für eine POS-Annotation (Stanford-NLP-Tagger, XML-Ausgabe)

```
<pos>
 <sentence id="0">
  <word wid="0" pos="NNP">Hey</word>
  <word wid="1" pos="NNP">Paul</word>
  <word wid="2" pos=".">!</word>
 </sentence>
 <sentence id="1">
  <word wid="0" pos="MD">Would</word>
  <word wid="1" pos="PRP">you</word>
  <word wid="2" pos="VB">give</word>
  <word wid="3" pos="PRP">me</word>
  <word wid="4" pos="DT">the</word>
  <word wid="5" pos="NN">hammer</word>
  <word wid="6" pos=".">?</word>
 </sentence>
</pos>
```

Annotation und Auszeichnungssprache

Annotation erfolgt mit Hilfe von Auszeichnungssprachen. Eine konkrete Auszeichnungssprache besteht aus folgenden Komponenten (*Tripod-Modell* nach Sperberg-McQueen und Huitfeldt 1999):

- Syntax zur Trennung von Auszeichnung und Primärdatum
- Formales Modell
- Dokumentgrammatik

Zur verbesserten Austauschbarkeit werden Auszeichnungssprachen auf Basis von Metasprachen entwickelt

XML als Basis linguistischer Annotation

Seit der Standardisierung der Metasprache SGML 1986 hat deren 1999 veröffentlichte Weiterentwicklung (zunächst als echte Teilmenge) XML die Basis für standardisierte linguistische Annotationsformate gelegt.
Als Metasprache stellt XML folgende Komponenten für Auszeichnungssprachen bereit:

- Die konkrete Syntax
- Das formale Modell
- Formalismen zur Erstellung von Dokumentgrammatiken

XML: Syntax

- Es gibt Elemente und Attribute
- Elemente bestehen aus Start- und Endtag, die den auszuzeichnenden Text (oder andere Elemente) umschließen
- Spitze Klammern „<“ und „>“ trennen Tag und Primärdatum
- Elemente können andere Elemente oder Text beinhalten
- Attribute sind den Elementen zugeordnet

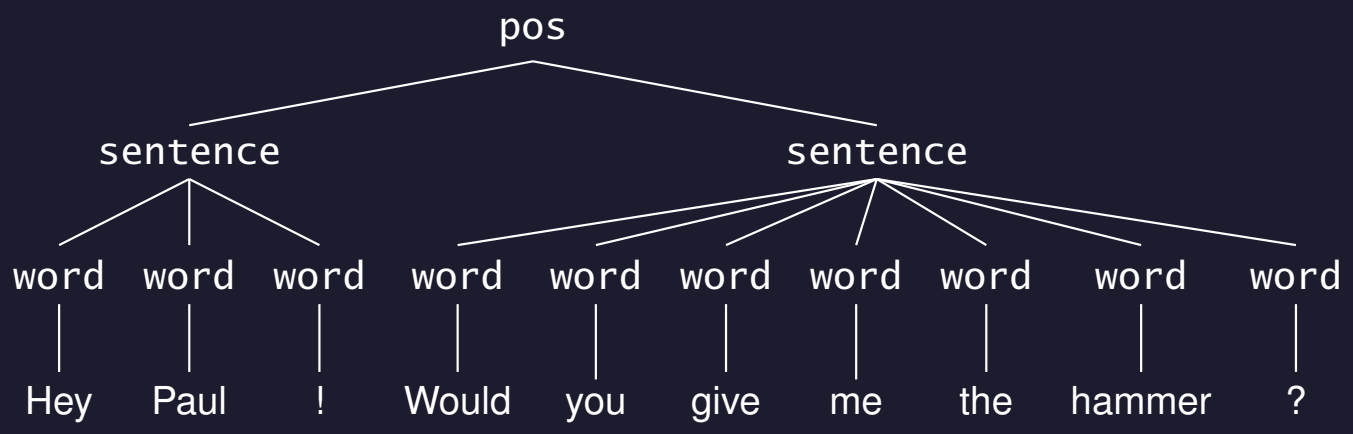
XML: Datenmodell

Das Datenmodell XML-basierter Auszeichnungssprachen folgt der Struktur von Texten

- „a text [is] an ‚ordered hierarchy of content objects‘“ (OHCO-These, DeRose u. a. 1990, S. 4)
- Eine solche Hierarchie wiederum kann als Baum mit einer einzelnen Wurzel dargestellt werden
- Das formale Modell von XML-Instanzen folgt der OHCO-These, d. h.
  - Es gibt genau ein Element (davor darf nur der so genannte Prolog stehen), das alle anderen Elemente beinhaltet (Wurzelelement)
  - Alle anderen Elemente, deren Start-Tag im Inhaltsmodell eines anderen Elements sind, haben auch das End-Tag im gleichen Inhaltsmodell, d. h. es gibt keine Überlappungen
- Diese Regeln formulieren die Qualität der *Wohlgeformtheit*, d. h. das entsprechende Dokument ist *wohlgeformt* – entspricht es darüber hinaus einer Dokumentgrammatik, ist es *gültig*

Datenmodell im Beispiel

Grafische Übersicht der Stanford-NLP-Taggerausgabe (ohne Attribute)



XML: Dokumentgrammatik

Eine Dokumentgrammatik legt folgende Eigenschaften einer Auszeichnungssprache fest:

- Anzahl und Art der Elemente, d. h. Namen der Elemente, Datentypen, etc.
- Verschachtelung der Elemente untereinander
- Art und Anzahl der Attribute
- Zuordnung der Attribute zu den Elementen

Im Umfeld von XML gibt es eine Reihe von Grammatikformalismen, mit denen Dokumentgrammatiken erstellt werden können

Einschränkung XML-basierter Auszeichnungssprachen

XML-basierte Auszeichnungssprachen haben aber nicht nur Vorteile:

- Darstellung multipler Annotationen problematisch
- Begrenzter semantischer Gehalt, d. h. einzig die Namen der Elemente und Attribute geben Hinweis auf die Motivation zur Auszeichnung

Multiple Annotation

Oftmals möchte man nicht nur eine Annotationsebene auszeichnen

Beispielkonstruktion

Hey Paul! Would you give me
the hammer?

Eigentlich sind noch weitere Informationen kodiert, die Sprecher:

Beispielkonstruktion

```
Peter: "Hey Paul! Would you give me"
Paul: "the hammer?"
```

Wir haben es hier mit zwei Äußerungen zu tun, die mit den Satzgrenzen überlappen

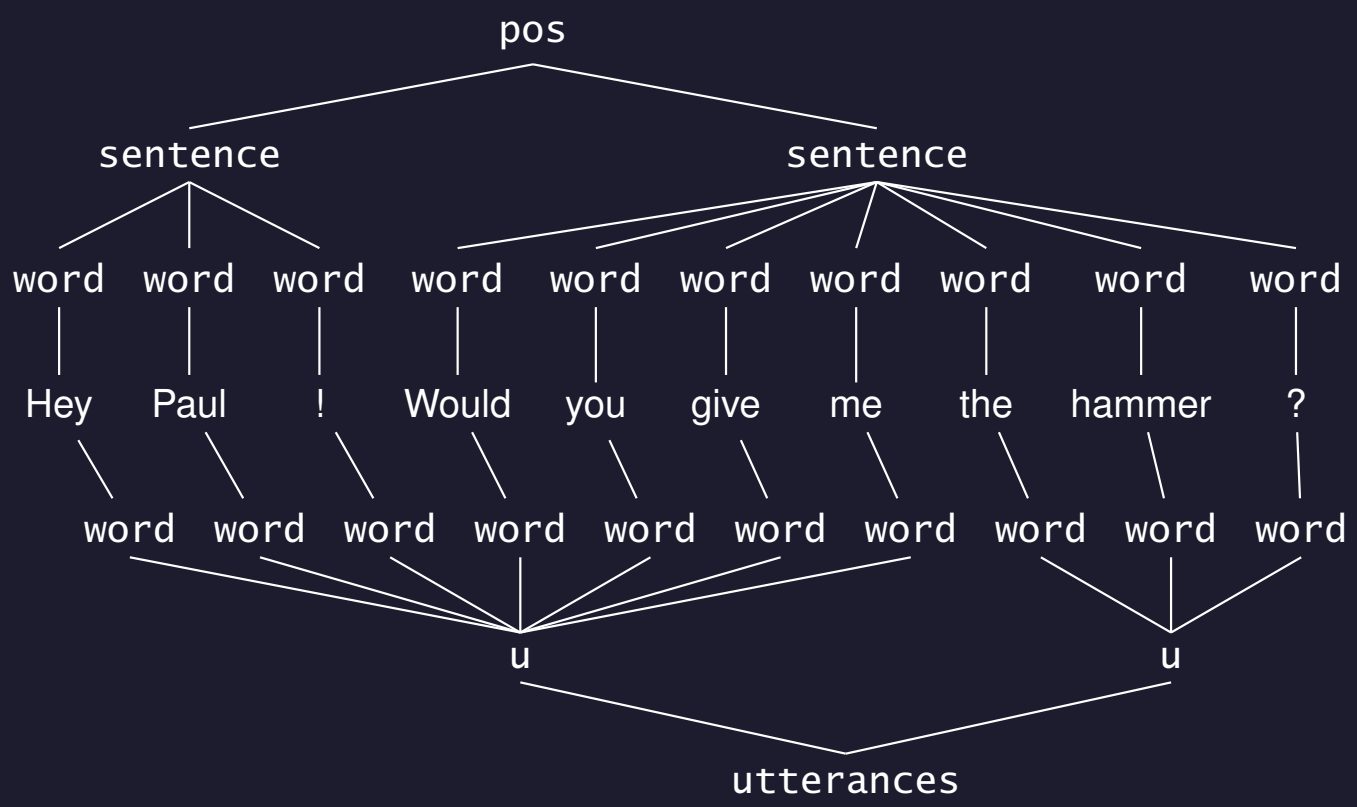
Multiple Annotation

Kombinierte Annotation

```
<pos>
 <sentence id="0">
  <u who="Peter">
   <word wid="0" pos="NNP">Hey</word>
   <word wid="1" pos="NNP">Paul</word>
   <word wid="2" pos=".">!</word>
 </sentence>
 <sentence id="1">
  <word wid="0" pos="MD">Would</word>
  <word wid="1" pos="PRP">you</word>
  <word wid="2" pos="VB">give</word>
  <word wid="3" pos="PRP">me</word>
 </u>
 <u who="Paul">
  <word wid="4" pos="DT">the</word>
  <word wid="5" pos="NN">hammer</word>
  <word wid="6" pos=".">?</word>
 </sentence>
</pos>
```

Multiple Annotation

Grafische Übersicht



Multiple Annotation

Problem: überlappende Annotationen sind in XML nicht repräsentierbar (vgl. Definition von Wohlgeformtheit)

- Überlappungen lassen sich definieren als „multiple parentage“ (Sperberg-McQueen und Huitfeldt 2004, S. 151), hier zwischen einzelnen `word`-Elementen und den Elementen `sentence` bzw. u – die Folge: Verletzung des formalen Modells
- Die übliche Vorgehensweise der Speicherung der Primärdaten und der beiden Annotationsebenen in einer Datei resultiert in einer nicht-wohlgeformten XML-Instanz

Lösungsansätze

Verschiedene Lösungsansätze zur Darstellung überlappender Annotationen bzw. diskontinuierlicher Annotationseinheiten werden diskutiert:

- Multiple Dokumente
- Meilensteine
- Fragmentierungen
- Standoff-Notation

Multiple Dokumente

Vorgehensweise: Speicherung jeweils einer Annotationsebene inkl. Primärdaten in einer Datei (teilweise auch inkl. einer grundlegenden Annotationsebene, vgl. Marinelli u. a. 2008)

Vorteile

- Jede Datei ist vollständig und einzeln verwendbar
- Gut verarbeitbar (sowohl durch menschliche als auch maschinelle Leser)
- Für jede Annotationsebene kann eine separate Dokumentgrammatik erstellt werden

Nachteile

- Primärdaten werden redundant gespeichert
- Bezug der Annotationen untereinander nur schwer und aufwändig zu realisieren
- Geringe Robustheit bzgl. Integrität der Primärdaten

Meilensteine

Vorgehensweise: Verwendung spezieller leerer Elemente, die jeweils den Beginn und das Ende einer Annotation definieren (TEI P2, Abschnitt 22.3.4.3, „Milestone method“)

Vorteile

- Ursprüngliche Elementinformationen und -typen können in Form von Attributen gespeichert werden
- Unterscheidung zwischen generischen Elementen und solchen mit eigenem semantischen Gehalt möglich (z. B. `lb` für einen Zeilenumbruch) und generischen Elementen (`milestone` und `anchor`)

Nachteile

- Bei vielfacher Überlappung unübersichtlich
- Besondere Formen von Überlappung (bsp. *Self Overlap*) nicht abbildbar
- Schlecht maschinell verarbeitbar
- Keine Unterscheidung von Inklusion und Dominanzbeziehungen

Fragmentierungen

Vorgehensweise: Aufbrechen der einzelnen überlappenden Elemente in kleinere Teilfragmente (*Partial Elements*, vgl. P5 2.0.2, Kapitel 20), die sich ohne Überlappungen in die Dokumentstruktur einbetten lassen

Vorteile

- Durch Hinzufügen des `part`-Attributs wird die logische Zusammengehörigkeit der Fragmente deutlich gemacht
- Self Overlap durch Verwendung des `next`-Attributs prinzipiell abbildbar

Nachteile

- Bei vielfacher Überlappung unübersichtlich
- Schlecht maschinell verarbeitbar
- Keine Unterscheidung von Inklusion und Dominanzbeziehungen

Standoff-Notation

Vorgehensweise: Trennung von Primärdaten und Markup und anschließende Referenzierung durch Zeigemechanismen

Vorteile

- Beliebig viele Annotationsebenen kombinierbar
- Prinzipiell gut skalierbar, da Verwendung von beliebiger Anzahl von Dateien möglich
- Verschiedene Serialisierungsformate vorstellbar – inkl. Unterscheidung von Inklusion und Dominanzbeziehung (vgl. Stührenberg und Jettka 2009)

Nachteile

- Für menschliche Leser sehr schlecht verarbeitbar
- Maschinelle Verarbeitung problematisch
- Je nach Ansatz geringe Robustheit bzgl. Integrität der Primärdaten

Standoff-Notation

Mögliche (nicht zwingend ideale) Realisierung

```
<root>
 <utterances start="0" end="39"/>
 <pos start="0" end="39"/>
 <u who="Peter" start="0" end="27"/>
 <u who="Paul" start="28" end="39"/>
 <sentence id="0" start="0" end="9"/>
 <sentence id="1" start="10" end="39"/>
</root>
```

- Neben der Verwendung von Zeichenpositionen sind auch andere Adressierungsmechanismen denkbar
- Standoff-Notation ermöglicht in Kombination mit dem XML-inhärenten Referenzierungsmechanismus (in Abhängigkeit einer Dokumentgrammatik) auch Graphen-basierte Datenmodelle (vgl. XStandoff ▸ Info)

Erweiterung des Komponenten-Modells

Erweiterung des „Tripod“-Modells zur Charakterisierung von Auszeichnungssprachen:

- Linearisierung
  - Syntax
  - Notation
- Datenmodell
- Validierungsmechanismus/Grammatik

XML und Semantik

- Die konkrete Syntax einer Auszeichnungssprache (sprich: die Trennung von Primärdatum und Auszeichnung) wird von der Metasprache vererbt
- Das Annotationsinventar wird durch die Dokumentgrammatik festgelegt

TEI-Annotation

```
<biblStruct>
 <monogr>
  <author>Charles F. Goldfarb</author>
  <title>The SGML Handbook</title>
  <edition>first edition</edition>
  <imprint>
   <publisher>Oxford University Press</publisher>
   <pubPlace>Oxford</pubPlace>
   <date>1991</date>
  </imprint>
 </monogr>
</biblStruct>
```

Nonsense-Annotation

```
<x>
 <y>
  <z>Charles F. Goldfarb</z>
  <w>The SGML Handbook</w>
  <g>first edition</g>
  <h>
   <a>Oxford University Press</a>
   <b>Oxford</b>
   <c>1991</c>
  </imprint>
 </y>
</x>
```

XML und Semantik

- Annotationsinventar gibt nur Hinweise auf die Motivation der Auszeichnung
- Weitere Semantik erfolgt durch Verknüpfung mit einer externen Ressource (z. B. Datenkategorienregistierung) – durch Verwendung generischer Auszeichnungselemente

NP-Annotation

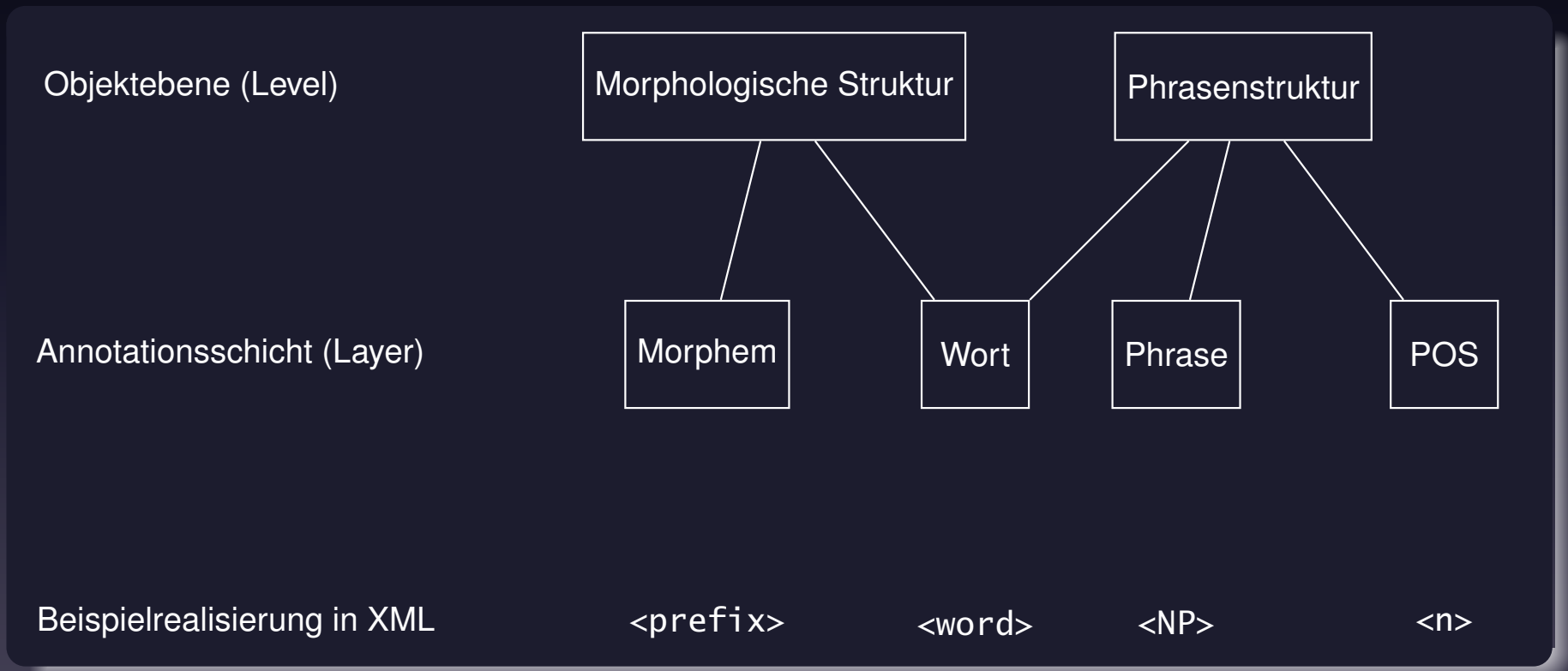
```
<NP NUM="sing" PER="third"/>
```

Feature Structure

```
<fs>
 <f name="CAT">
  <symbol value="np" dcr="#datcat/DC-1333"/>
 </f>
 <f name="AGR">
  <fs>
   <f name="NUM">
    <symbol value="sing" dcr="#datcat/DC-1387"/>
   </f>
   <f name="PER" />
    <symbol value="third" dcr="#datcat/DC-1402"/>
   </f>
  </fs>
 </f>
</fs>
```

XML und Semantik

Aufbauend auf generischer Auszeichnung und Standoff-Notation lässt sich eine weitere Komponenten von Auszeichnungssprachen identifizieren: die Trennung von Konzept und Serialisierung (eingeführt von Bayerl u. a. 2003; Witt 2004)



Vollständiges Komponenten-Modell

Finale Version des Komponenten-Modells zur Charakterisierung von Auszeichnungssprachen:

- Linearisierung
  - Syntax
  - Notation
- Datenmodell
- Validierungsmechanismus/Grammatik
- Unterscheidung von konzeptueller Ebene und Serialisierung

Gruppen von Auszeichnungssprachen

Auf Basis des erweiterten Komponenten-Modells lassen sich Auszeichnungssprachen in verschiedene Gruppen einordnen

Syntax

XML vs nicht-XML

Notation

Inline vs. Standoff

Datenmodell

Baum vs. Graph

Dokumentgrammatik

Merkmale und formale Ausdrucksstärke

Unterscheidung von Konzept und Serialisierung

Unterstützt vs. nicht-unterstützt

Gliederung

Relavante Standards

De-facto-Standards:

- Guidelines der Text Encoding Initiative (TEI Guidelines)
- Corpus Encoding Standard (XML-Version: XCES)

De-jure-Standards:

- Linguistic Annotation Framework (LAF/GrAF, ISO/FDIS 24612)
- Morpho-Syntactic Annotation Framework (MAF, ISO/DIS 24611)
- Syntactic Annotation Framework (SynAF, ISO 24615:2010)
- Data Category Registry (DCR, ISO 12620:2009)

Darüber hinaus die standardisierte Beschreibung von Merkmalsstrukturen (als Teil der TEI Guidelines und ISO 24610-1:2006; ISO 24610-2:2011)

TEI Guidelines

Die Guidelines der Text Encoding Initiative (TEI Guidelines)

- Erste (nicht-öffentliche) Version (P1) 1987, aktuell: P5 2.0.2
- Modular aufgebautes Tagset zur Strukturierung verschiedenster geisteswissenschaftlicher Texte
- bis zur TEI P3 SGML-basiert, ab der P4 XML-basiert
- über 1600 Seiten starken Dokumentation
- Entwickelt durch ein internationales Konsortium aus Wissenschaftlern, Bibliothekaren und weiteren Personen und Organisationen
- Theoriefrei
- Open Source

TEI Guidelines

Das TEI-Tagset enthält neben dem obligatorischen Modulen TEI, Core und Header (Metadaten) u. a. Module für...

- Verse
- Aufführungstexte
- Transkriptionen
- Wörterbücher
- Manuskripte
- Analoge Primärquellen
- Anmerkungen
- Namen, Daten, Personen, Orte
- Tabellen, Formeln, Grafiken, Musik
- Sprachkorpora
- Merkmalsstrukturen
- Graphen
- ...

TEI-Ausschnitt

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
  <!-- [...] -->
  <interpretation>
   <p>POS-Analyse auf Basis des Stanford-NLP-Taggers</p>
  </interpretation>
  <!-- [...] -->
 </teiHeader>
 <text>
  <body>
   <p>
    <s xml:id="s1">
     <w xml:id="s1w1" ana="#NNP">Hey</w>
     <w xml:id="s1w2" ana="#NNP">Paul</w>
     <w xml:id="s1w3" ana="#.">!</w>
    </s>
    <s xml:id="s2">
     <w xml:id="s2w1" ana="#MD">Would</w>
     <w xml:id="s2w2" ana="#PRP">you</w>
     <!-- [...] -->
    </s>
   </p>
  </body>
 </text>
</TEI>
```

Bewertung

- Die TEI Guidelines sind seit über 20 Jahren weltweit im Einsatz
- Neue Fassungen sind entweder abwärtskompatibel oder erlauben die Transformation bestehender Daten
- Für die Annotation linguistischer Daten bietet die TEI Elemente zur Segmentierung hinunter auf Zeichenenebene
- Die TEI unterstützt eine Vielzahl an Mechanismen zur Speicherung multipler Annotationen
- Für die Auszeichnung konkreter linguistischer Merkmale fehlen Elemente und Attribute – hier müssen externe Ressourcen herangezogen werden
- Das Tagset ist sehr komplex und ermöglicht mehrere Wege, um Phänomene zu beschreiben, daher zusätzliche Annotation Guidelines notwendig

XCES

- Im Rahmen des EU-Projekts *Expert Advisory Group on Language Engineering Standards* (EAGLES) wurden die *EAGLES Guidelines* erarbeitet, eine Ansammlung an Empfehlungen für De-facto-Standards und *Good Practice* im Bereich der Sprachverarbeitung, angelehnt an die TEI Guidelines
- Teil besagter Guidelines ist der Corpus Encoding Standard (CES), ein Anwendung der TEI Guidelines P3 (in Form einer Modifikation)
- Die XML-Fassung XCES (vgl. Ide u. a. 2000) unterscheidet sich als nicht zur TEI P4 kompatible Weiterentwicklung teilweise deutlich von der ursprünglichen Fassung
- Beim IDS in Mannheim wird aktuell an einer TEI P5-kompatiblen XCES-Version gearbeitet (IDS-XCES)

XCES: Aufbau

- XCES sieht eine Kodierung der Primärdaten (in Form einer Basisannotation) sowie die Annotation der eigentlichen linguistischen Phänomene vor
- Die Strukturierung der Primärdaten wird in der Dokumentgrammatik `cesDoc` definiert:
  - Level 1 (minimale Segmentierung): Kennzeichnung der logischen Dokumentstruktur bis auf Absatzebene; Hervorhebungen durch das generische Element `hi`; Metadaten im CES-Header
  - Level 2: Ersetzung des generischen `hi`-Elements durch spezifischere Annotation; einheitliche Annotation eines gegebenen linguistischen Phänomens
  - Level 3: Abkürzungen, Nummern, Fremdwörter und -Phrasen sind entsprechend ausgezeichnet; Morphosyntaktische Annotation werden durch ein vom Benutzer vorgegebenes Tagset annotiert; Sätze und wörtliche Rede sind durch die entsprechenden Elemente `s` und `q` ausgezeichnet.
- Metadaten (analog zur TEI) werden in einer externen Datei gespeichert
- Annotationsebenen werden mittels Merkmalsstrukturen (definiert in der Dokumentgrammatik `cesAna`) kodiert; die eigentliche Annotation erfolgt in Standoff-Notation auf Basis der Primärdatenkodierung

XCES: Beispiel-Kodierung der Primärdaten

Level-1-konforme Primärdaten

```
<?xml version="1.0" encoding="UTF-8"?>
<cesDoc xmlns="http://www.xces.org/schema/2003" version="0.4">
 <text>
  <body>
   <p>The Story Continues . . . a serial enovel by Ferd Eggan</p>
   <p>1 Welcome to Hotel Real Desert</p>
   <p>But he never fell into the error of arresting his intellectual development by any formal
    acceptance of creed or system, or of mistaking, for a house in which to live, an inn that is but
    suitable for the sojourn of a night in which there are no stars and the moon is in travail</p>
   <p>The Hotel</p>
   <p>Hotel is next door to a perfect metaphor for the mind, and thus for psychoanalysis. In my
    father's house are many mansions?To get there you have to leave somewhere else...</p>
   <!-- [...] -->
  </body>
 </text>
</cesDoc>
```

XCES: Metadaten

XCES-Metadaten (OANC-Beispiel)

```
<?xml version="1.0" encoding="UTF-8"?>
<cesHeader creator="KBS" date.created="20050222">
 <fileDesc>
  <titleStmt>
   <title>The Story Continues</title>
   <author>Ferd Eggan</author>
  </titleStmt>
  <sourceDesc><!-- [...] --></sourceDesc>
 </fileDesc>
 <profileDesc>
  <textClass>
   <domain>Fiction</domain>
   <subdomain>General fiction</subdomain>
   <!-- [...] -->
  </textClass>
  <annotations>
   <annotation ann.loc="TheStory.txt" type="content">Text content</annotation>
   <annotation ann.loc="TheStory-logical.xml" type="logical">Logical structure</annotation>
   <!-- [...] -->
  </annotations>
 </profileDesc>
</cesHeader>
```

XCES: Annotationsebenen

XCES-Annotation (OANC-Beispiel)

```
<?xml version="1.0" encoding="UTF-8"?>
<cesAna xmlns="http://www.xces.org/schema/2003" version="1.0.4">
 <struct type="cesDoc" from="0" to="400307">
  <feat name="xmlns" value="http://www.xces.org/schema/2003"/>
  <feat name="version" value="1.0.4"/>
 </struct>
 <struct type="text" from="2" to="400306"/>
 <struct type="body" from="5" to="400304"/>
 <struct type="div" from="9" to="73"/>
 <struct type="p" from="14" to="69">
  <feat name="id" value="p1"/>
 </struct>
 <struct type="head" from="77" to="108">
  <feat name="type" value="h1"/>
 </struct>
 <struct type="p" from="112" to="414">
  <feat name="id" value="p2"/>
 </struct>
 <struct type="hi" from="409" to="410">
  <feat name="rend" value="sup"/>
 </struct>
 <!-- [...] -->
</cesAna>
```

Bewertung

- XCES ist auf Grund der eigenständigen Entwicklung aktuell nicht kompatibel zur TEI P5
- Dokumentation ist nur unzureichend vorhanden (teilweise wird auf CES-Elemente verwiesen, die in XCES nicht vorhanden sind)
- Die Dokumentgrammatik in Form mehrerer XML Schemata ist kaum dokumentiert und unzureichend versioniert
- Zwar steht die zweite Fassung des OANC (*Second Release*) XCES-kodiert zum Download zur Verfügung, aktuell wird der OANC allerdings in LAF/GrAF rekodiert

Feature Structures

Feature Structures (Merkmalsstrukturen, vgl. Carpenter 1992; Copestake 2002) beschreiben allgemein die Eigenschaften eines Objekts in einer abstrakten hierarchisch organisierten Form und sind in der Linguistik weit verbreitet:

- In der *Head-driven Phrase Structure Grammar* (HPSG, Pollard und Sag 1987; Pollard und Sag 1994)
- In der *Lexical Functional Grammar* (LFG, Kaplan und Bresnan 1982; Dalrymple 2001) zur Darstellung syntaktischer Kategorien in der F-Struktur

Ein wesentlicher Vorteil ist die Möglichkeit der Unifikation von Merkmalsstrukturen

Eine einfache Merkmalsstruktur

person	
vorname	max
nachname	meier
geschlecht	männlich

TEI Feature Structures

- Die TEI Guidelines beinhalten bereits seit langer Zeit Ausführungen zur Beschreibung von Merkmalsstrukturen
- Bei der Erstellung der TEI P5 wurde das Kapitel entsprechend überarbeitet, die aktuelle Fassung unterscheidet sich teilweise deutlich von der in CES/XCES genutzten Serialisierung
- Gleichzeitig wurden die Ausführungen im Rahmen der internationalen Normen ISO 24610-1:2006 und ISO 24610-2:2011 standardisiert
- ISO 24610-1:2006 beschreibt dabei die Merkmalsstrukturen als solche, während ISO 24610-2:2011 Merkmalssystembeschreibungen (*Feature System Declaration*) beschreibt
- Aktuell sind die TEI Feature Structures kompatibel zu den genannten Normen (mit Ausnahme unterschiedlicher Wurzelelemente und weniger zusätzlicher Ausführungen formaler Art in den Normendokumenten)

TEI Feature Structures

TEI-Repräsentation der Merkmalsstruktur

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader>
  <!-- [...] -->
 </teiHeader>
 <text>
  <body>
   <p></p>
   <fs type="person">
    <f name="vorname">
     <string>max</string>
    </f>
    <f name="nachname">
     <string>meier</string>
    </f>
    <f name="geschlecht">
     <symbol value="männlich"/>
    </f>
   </fs>
  </body>
 </text>
</TEI>
```

Bewertung

Bis auf wenige Unterschiede sind die Ausführungen der TEI Guidelines inhaltsgleich mit den Normen, so dass die Auswertung für beide Varianten gilt

- Die standardisierte Beschreibung von Merkmalsstrukturen erlaubt die Trennung von Konzept und Serialisierung auf Grund der generischen Elemente und Attribute
- Effiziente Beschreibung vorrangig datenzentrierter Informationen (z. B. Lexika)
- Verwendung des formalen Modells des gerichteten, einwurzeligen azyklischen Graphen (DAG) – damit auch Abbildung multipler Annotationen möglich
- Die Dokumentgrammatik macht keinerlei Restriktionen bzgl. der Benennung von Merkmalen und Werten – der Einsatz einer externen Datenkategorienregistrierung ist notwendig

Bewertung

Dennoch gibt es auch Unterschiede zwischen beiden Ausführungen:

- Die TEI Guidelines wurden nach Verabschiedung der Norm ISO 24610-1:2006 mehrfach geändert (zuletzt 2012, P5 2.0.2), diese Aktualisierungen fehlen im Standard zu Beschreibung einer Merkmalsstruktur
- Die Norm zur Serialisierung von Merkmalssystembeschreibungen ISO 24610-2:2011 wurde erst 2011 veröffentlicht, es muss davon ausgegangen werden, dass sie nicht mehr vollständig kompatibel zu ISO 24610-1:2006 ist
- Die für ISO 24610-1:2006 fällige Überarbeitung nach Auslaufen einer ISO-üblichen Fünfjahresfrist blieb aus, seit Anfang 2012 ist der Standard auf den Status „suspended“ gesetzt

Linguistic Annotation Framework (LAF)

Das Linguistic Annotation Framework (LAF) stellt ein generalisiertes Konstrukt zur Darstellung von (multiplen) Annotationen dar

- Entwicklungsbeginn 2002, aktuell im Status eines Final Draft International Standard, finale Fassung voraussichtlich 2012
- LAF gibt kein Tagset vor, sondern definiert
  - Ein abstraktes Datenmodell (*Data Model*)
  - Die Segmentierung (*Base Segmentation*)
  - Mögliche Serialisierungsformate (zu unterscheiden zwischen der Serialisierung in der vom Anwender definierten Dokumentenformat, *Document Form*, und dem Austauschformat (*Pivot Format* oder auch *Dump Format*) GrAF, das *Graph Annotation Format*
- Die Norm beansprucht, unabhängig von Medientypen und formal eindeutig zu sein
- LAF und GrAF nehmen dabei Anleihen an vorgehenden Entwicklungen wie CES/XCES

LAF: Aufbau

Analog zu CES/XCES werden Primärdaten, Annotation und Metadaten in separaten Dateien gespeichert und in Standoff-Notation annotiert

- Primärdaten liegen üblicherweise unannotiert vor
- In einer früheren Fassung vorgesehene unterschiedliche Header für Primärdaten (*Primary Data Document Header*) und Annotationsebenen (*Annotation Document Header*) wurden in der letzten Version zusammengefasst, auch sind in ISO/FDIS 24612 Metadaten und Annotation in einer Instanz gespeichert
- Die eigentliche Annotation wird mittels Merkmalsstrukturen kodiert, dabei weicht GrAF von der in ISO 24610-1:2006 bzw. TEI P5 2.0.2 vorgeschlagenen Serialisierung ab
- Bei der Verwendung von Typen- und Merkmalsdefinitionen ist die Einbindung einer externen Ressource notwendig

LAF und GrAF

Eine GrAF-Instanz

```
<?xml version="1.0" encoding="UTF-8"?>
<graph xmlns="http://www.xces.org/ns/GrAF/1.0/">
 <graphHeader><!-- [...] --></graphHeader>
 <node xml:id="fn-n156"/>
 <a label="FE" ref="fn-n156">
  <fs>
   <f name="FE" value="Speaker"/>
   <f name="rank" value="1"/>
   <f name="GF" value="Ext"/>
   <f name="PT" value="NP"/>
  </fs>
 </a>
 <!-- [...] -->
 <edge xml:id="e233" from="fn-n156" to="fn-n133"/>
 <!-- [...] -->
 <region xml:id="r1" anchors="980 9190"/>
 <!-- [...] -->
 <node xml:id="a232">
  <link targets="r1"/>
 </node>
 <!-- [...] -->
 <a label="R Gesture Units 1" ref="a232"/>
</graph>
```

Bewertung

- Problematisch ist, dass LAF kein Tagset vorgibt, lediglich GrAF ist in Teilen in der aktuellen Version der Spezifikation definiert
- Da GrAF nur als intermediäres Austauschformat fungiert, sind prinzipiell unendlich viele Annotationsformate mit LAF kompatibel (sofern das Graphen-basierte Datenformat darauf abbildbar ist)
- In der Entwicklung von LAF hat es viele unterschiedliche Fassungen des Datenmodells und des Serialisierungsformats gegeben, wissenschaftliche Publikationen vor 2011 sind daher nicht mehr aktuell
- Die Verwendung einer eigenen Serialisierung für Merkmalsstrukturbeschreibungen ist zumindest als unglücklich zu bezeichnen, allerdings verweist LAF für komplexere Feature Structures auf ISO 24610-1:2006

Morpho-Syntactic Annotation Framework (MAF)

Da sowohl XCES als auch LAF konkrete Annotationen (morpho-)syntaktischer Phänomene ausklammern, gibt es mit ISO/DIS 24611 und ISO 24615:2010 zwei eng verwandte Normenvorschläge zu diesem Aspekt.

- Das Morpho-Syntactic Annotation Framework (MAF) definiert ein Metamodell zur Repräsentation morphosyntaktischer Annotationen sowie eine Datenkategorieauswahl (Data Category Sets, DCS)
- Ein annotiertes Dokument besteht aus den Primärdaten (*Raw Document*) und den eigentlichen Annotationen (im Standoff-Verfahren)
- Die Annotationsebene besteht aus Wortformen, die über Zeichenspannen der Primärdaten begrenzt werden; eine Wortform kann dabei einen Bereich über kein Segment oder mehrere Segmente umfassen und auf Lexikoneinträge verweisen
- Die Segmentierung erfolgt mittels Token, die – ebenso wie Wortformen – in gerichteten azyklischen Graphen (DAG) organisiert werden können
- Morphosyntaktische Informationen der Wortform werden mittels Merkmalsstrukturbeschreibungen gemäß ISO 24610-1:2006 annotiert
- Die Kategorien werden in einer Datenkategorienregistrierung gemäß ISO 12620:2009 gespeichert

MAF: Aufbau

- Token dienen als Anker für spätere Annotationen und definieren nicht-leere, fortlaufende Zeichenkettensegmente in den Primärdaten
- Zulässig ist die Verwendung mehrerer Token, die dieselbe Textspanne (oder überlappende Textspannen) markieren (z. B. bei Agglutinationen)
- Mehrere Wortformen können sich auf ein und dasselbe Token beziehen, so dass sich Überlappung auch auf dieser Ebene manifestieren lassen (ISO/DIS 24611, S. 16)
- MAF selbst definiert kein verbindliches Serialisierungsformat und verweist auf externe Metadaten-Standards
- Token und Primärdaten können inline- oder standoff-annotiert vorliegen (Standoff-Notation wird präferiert)

MAF: Segmentierung

Komposita-Darstellung (Inline-Notation)

```
<!-- [...] -->
<token form="Geburtstag" xml:id="t91" join="right">Geburtstags</token>
<token form="Geschenk" xml:id="t92" join="right">geschenk</token>
<token form="Papier" xml:id="t93">papier</token>
<wordForm tokens="#t91 #t92 #t93">
 <wordForm entry="urn:lexicon:de:geburstag" lemma="geburtstag" tokens="#t91"/>
 <wordForm entry="urn:lexicon:de:geschenk" lemma="geschenk" tokens="#t92"/>
 <wordForm entry="urn:lexicon:de:papier" lemma="papier" tokens="#t93"/>
</wordForm>
<!-- [...] -->
```

MAF: Beispiel-Instanz

<tiger2/>-XML als potentielles Serialisierungsformat (Standoff-Notation)

```
<?xml version="1.0" encoding="UTF-8"?>
<maf>
 <token id="t1" form="I" from="0" to="1"/>
 <token id="t2" join="right" form="wan" from="2" to="5"/>
 <token id="t3" join="left" form="na" from="5" to="7"/>
 <!-- [...] -->
 <wordForm lemma="I" tokens="t1"><!-- [...] --></wordForm>
 <wordForm lemma="want" tokens="t2">
  <fs>
   <f name="pos">
    <symbol value="VBP"/>
   </f>
  </fs>
 </wordForm>
 <wordForm lemma="to" tokens="t3">
  <fs>
   <f name="pos">
    <symbol value="TO"/>
   </f>
  </fs>
 </wordForm>
 <wordForm tokens="t2 t3"/>
 <!-- [...] -->
</maf>
```

Metadaten in MAF

OLAC-Metadaten zur Segmentierung

```
<?xml version="1.0" encoding="UTF-8"?>
<maf document="interview.mpeg" addressing="mpeg7">
 <olac:olac xmlns:olac="http://www.language-archives.org/OLAC/1.0/"
  xmlns="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.org/dc/terms/">
  <creator>Maik Stührenberg</creator>
  <date>2011-12-14</date>
  <description>Händisch angepasstes MAF-Beispiel auf Basis des in ISO/DIS 24611, S. 35 enthaltenen Fragments
       .</description>
 </olac:olac>
 <token id="t0" from="T00:01:16:4484F30000" to="T00:01:16:14494F30000"
  transcription="mister"/>
 <wordForm tokens="t0" lemma="mister">
  <!-- ... -->
 </wordForm>
 <!-- ... -->
</maf>
```

Bewertung

- MAF befindet sich aktuell im Ballot zum FDIS, es ist abzuwarten, welche Änderungen bis zur finalen Fassung vorgenommen werden
- Die Spezifikation bindet sinnvoll andere vorhandene ISO-Normen ein und vermeidet Dopplungen
- Auch wenn MAF selbst kein Serialisierungsformat definiert, steht mit dem <tiger2/>-XML-Format ein geeigneter Kandidat zur Verfügung (Anfang Juni 2012 als Standardisierungsvorschlag 24615-2 eingereicht)

Syntactic Annotation Framework (SynAF)

Das 2010 als Internationale Norm ISO 24615:2010 verabschiedete Syntactic Annotation Framework (SynAF) kann als Ergänzung zu MAF angesehen werden

- Der Standard definiert ein Metamodell zur syntaktischen Annotation in Verbindung mit Datenkategorien in Form eines Data Category Sets (DCS)
- Im Gegensatz zu MAF, das die Annotation von POS-, morphologischen und grammatikalischen Informationen zum Ziel hat, können mit Hilfe von SynAF Annotationen der syntaktischen Konstituenten (in Form von Gruppen morphosyntaktisch annotierter Einheiten) innerhalb der Satzgrenzen standardisiert vorgenommen werden (Declerck 2006, S. 229)
- SynAF unterstützt dabei sowohl die Abbildung linguistischer Konstituenten (wie z. B. NPs) als auch von Dependenzstrukturen (ISO/FDIS 24615, S. 9)

Data Category Registry (DCR)

Eine gesicherte Erkenntnis ist, dass es nicht *das eine* Auszeichnungsinventar gibt – daher ist eine zukünftige Strategie:

- Bereitstellung eines generischen Annotationsmechanismus'
- Schaffung einer zentralen Kategorisierungsstelle zur standardisierten Verwaltung von Annotationsmerkmalen (also den Merkmalen und deren konkreten Werten) – *Data Category Registry* (DCR, etwa „Datenkategorienregistrierung")
- Verwendung standardisierter Datenkategorien (DCS)

Die internationale Norm ISO 12620:2009 beschreibt ein Datenmodell und Verfahren zur Verwaltung von Datenkategorien

Data Category Registry (DCR)

Ein Beispiel

```
<text>
<para>Ein Absatz</para>
</text>
```

Alternative Repräsentation

```
<text>
<p>Ein Absatz</p>
</text>
```

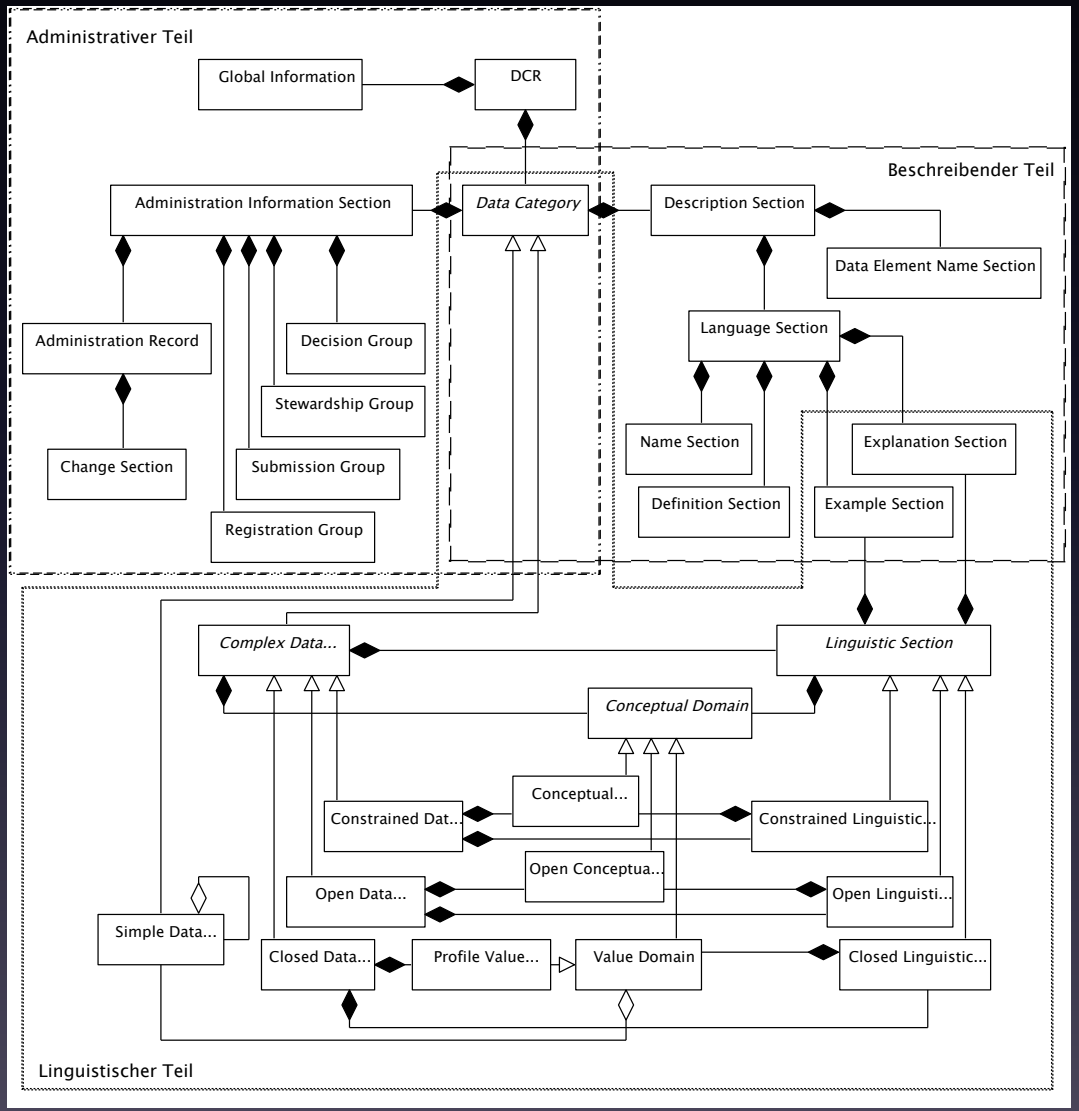
- Die Elemente `para` und `p` unterscheiden sich in ihrem *Generic Identifier*, bezeichnen aber beide das Konzept eines Absatzes in der logischen Dokumentstruktur
- Für menschliche Leser ist die Abbildung aufeinander trivial, nicht unbedingt für Maschinen – außer beide Elemente verweisen auf die selbe Datenkategorie „Absatz in der logischen Dokumentstruktur“ verweisen (z. B. repräsentiert durch einen URI)
- Datenkategorien ergänzen damit die Syntax einer Auszeichnungssprache um eine semantische Komponente und adressieren damit einen wesentlichen Aspekt einer gemeinsamen Sprache: „Real interoperability is a function of shared semantics, not syntax.“ (Bray 2005, S. 3)

DCR: Aufbau

Eine Datenkategorie im Sinne von ISO 12620:2009 besteht aus drei Bereichen:

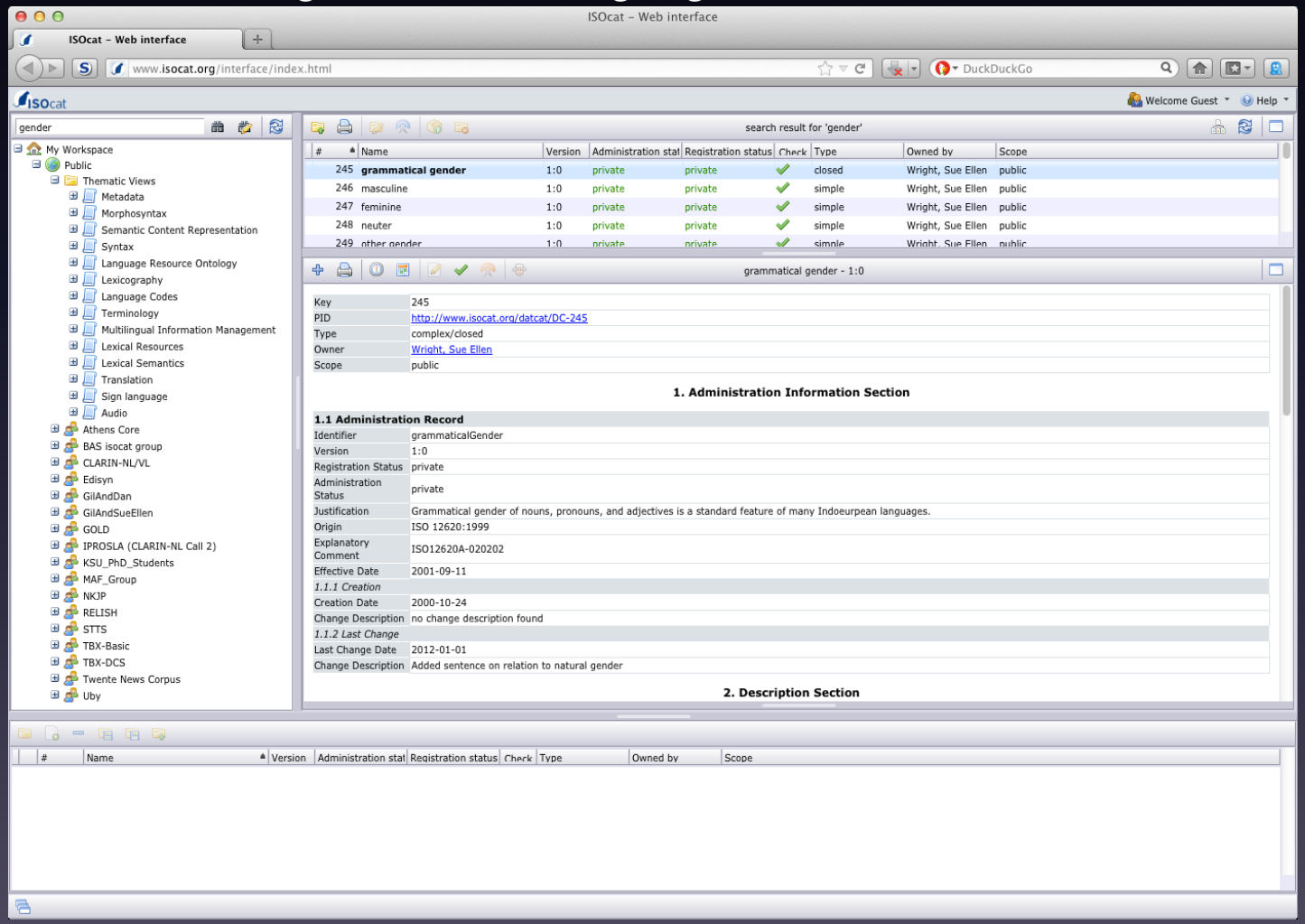
- Administrative Informationen
- Beschreibende Informationen
- Linguistische (bzw. sprachbezogene) Informationen

DCR: Aufbau


Administrativer Teil
Global Information
DCR
Beschreibender Teil
Administration Information Section
Data Category
Description Section
Data Element Name Section
Language Section
Administration Record
Decision Group
Stewardship Group
Change Section
Submission Group
Registration Group
Name Section
Explanation Section
Definition Section
Example Section
Complex Data...
Linguistic Section
Conceptual Domain
Conceptual...
Constrained Dat...
Constrained Linguistic...
Open Conceptua...
Open Data...
Open Linguisti...
Simple Data...
Closed Data...
Profile Value...
Value Domain
Closed Linguistic...
Linguistischer Teil

DCR: Implementierung

Mit ISOcat steht eine web-basierte Oberfläche für die Erstellung und das Retrieval von Datenkategorien zur Verfügung:



DCR: Implementierung

In ISOcat sind bereits eine Reihe von Datenkategorien enthalten, die für die eigene Annotation verwendet werden können

- CLARIN-Projektverbund
- GOLD (General Ontology for Linguistic Description)
- STTS
- ...

Verknüpfung von Annotation und Datenkategorie erfolgt über weltweit eindeutige URI

Bewertung

- Die internationale Norm ISO 12620:2009 und vor allem deren Implementierung ISOcat ist ein äußerst nützliches Werkzeug zur Anreicherung von Auszeichnungen mit erweitertem semantischen Gehalt
- Unter `http://www.isocat.org/` steht neben der Plattform selbst auch umfangreiche Dokumentation zur Verfügung

Gliederung

Bestandsaufnahme

Die gute Nachricht:

Linguistische Annotation wird als ernsthaftes Betätigungsfeld für Normung angesehen

Grundlegende Standards

XML als Metasprache ist in der linguistischen Annotation fest etabliert

Bestandsaufnahme

Internationale Normen setzen verstärkt auf generische Annotationsformate in Kombination mit standardisierten Datenkategorien – das muss nicht zwingend die beste Wahl für ein gegebenes Projekt sein:

Bański und Przepiórkowski (2010, S. 100)

„A tendency may be observed of increasing abstractness and generality of proposed standards [...]. This leads to their greater formal elegance, at the cost of their actual usefulness.“

Vergleich De-facto- und De-jure-Standards

Die TEI Guidelines haben großen Einfluss auf die aktuellen Standardisierungsbemühungen:

- Die P3 war Grundlage für den Corpus Encoding Standard (CES)
- Das in XCES formalisierte Generic Mapping Tool (GMT) kann als Vorläufer der standardisierten Beschreibung von Merkmalsstrukturen angesehen werden
- Die in der aktuellen TEI P5 enthaltene standardisierte Beschreibung von Merkmalsstrukturen ist in ISO 24610-1:2006 und ISO 24610-2:2011 normiert

Dagegen haben internationale Normen potentielle Nachteile:

- Standardisierten Auszeichnungssprachen (im Sinne von internationalen Normen) fehlen oftmals konkrete Dokumentgrammatiken
- Normen sind nicht immer kostenfrei verfügbar – oftmals werden in der Wissenschaft Vorversionen genutzt, deren Inhalt aber noch weit von der finalen Fassung abweichen kann

Handlungsanweisungen

Für den Aufbau eines Korpus

- Es gibt eine ganze Reihe an De-facto- und De-jure-Standards, die für die linguistische Annotation genutzt werden können – bitte keine weiteren Auszeichnungssprachen entwickeln!
- Dokumentation, Dokumentation, Dokumentation! (vgl. Negativbeispiel CES/XCES)
- Metadaten, Metadaten, Metadaten!

Handlungsanweisungen: Notation

Standoff-Notation wird als Gegengewicht zur Inline-Annotation ausgebaut – vor allem bei Mehr-Ebenen-Auszeichnung

- Annotatoren und Wissenschaftler sollten sich *jetzt* mit den Prinzipien und Besonderheiten der Standoff-Notation vertraut machen
- Aktuell existiert nur wenig Software, die mit Standoff-annotierten Daten umgehen kann: Serengeti (Diewald u. a. 2008; Stührenberg, Goecke u. a. 2007) oder Glozz beispielsweise(Mathet und Widlöcher 2011; Widlöcher und Mathet 2009)

Handlungsanweisungen: Trennung von Konzept und Serialisierung

Verbunden mit der Umstellung von Inline- auf Standoff-Notation lässt sich festhalten:

- Standoff-Notation erleichtert bereits die Trennung von Konzept und Serialisierung
- Der Einsatz generischer Annotationsformate wie ISO 24610-1:2006 bzw. ISO/FDIS 24612 forciert die Trennung
- ISO 12620:2009 und dessen Implementierung ISOcat stellen ein standardisiertes Verfahren zur Beschreibung und Referenzierung von Konzepten zur Verfügung

Danke für die Aufmerksamkeit!

XStandoff

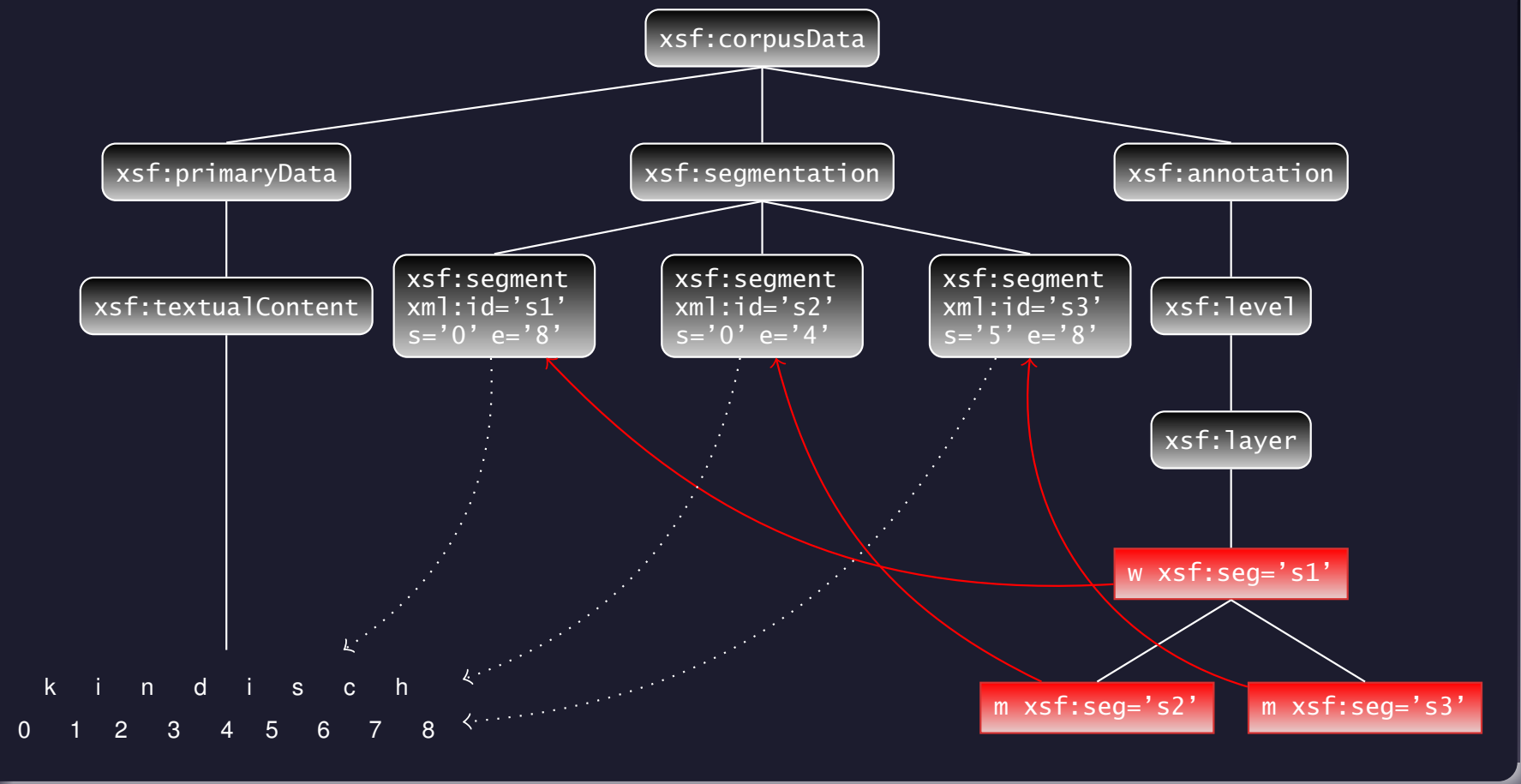
Alternativer Standoff-Ansatz:

XStandoff (vereinfachte Darstellung)

```
<?xml version="1.0" encoding="UTF-8"?>
<xsf:corpusData>
 <xsf:segmentation>
  <xsf:segment xml:id="s1" s="0" e="8"/>
  <xsf:segment xml:id="s2" s="0" e="4"/>
  <xsf:segment xml:id="s3" s="5" e="8"/>
 </xsf:segmentation>
 <xsf:annotation>
  <xsf:level>
   <xsf:layer>
    <w xsf:seg="s1">
     <m xsf:seg="s2" />
     <m xsf:seg="s3" />
    </w>
   </xsf:layer>
  </xsf:level>
 </xsf:annotation>
</xsf:corpusData>
```

XStandoff

Graphische Darstellung



Literatur I

Bański, Piotr und Adam Przepiórkowski (Juli 2010). „TEI P5 as a Text Encoding Standard for Multilevel Corpus Annotation“. In: *Digital Humanities 2010 Conference Abstracts*. The Alliance of Digital Humanities Organisations u. a. London, S. 98–100. url: http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-616.pdf (besucht am 05. 03. 2012).

Bayerl, Petra Saskia, Harald Lüngen, Daniela Goecke, Andreas Witt und Daniel Naber (2003). „Methods for the semantic analysis of document markup“. In: *Proceedings of the 2003 ACM symposium on Document engineering (DocEng)*. Hrsg. von Cecile Roisin, Ethan Muson und Christine Vanoirbeek. Grenoble: ACM Press, S. 161–170. doi: 10.1145/958220.958250.

Bray, Tim (Nov. 2005). „On Language Creation“. In: *Proceedings of the XML 2005 Conference*. Atlanta.

Burnard, Lou und Syd Bauman, Hrsg. (Feb. 2012). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.0.2. Last updated on 2nd February 2012. Oxford u. a.: Published for the TEI Consortium by Humanities Computing Unit, University of Oxford.

Carpenter, Bob (1992). *The Logic of Typed Feature Structures*. Cambridge: Cambridge University Press.

Copestake, Ann (2002). *Implementing Typed Feature Structure Grammars*. Stanford: CSLI Publications.

Dalrymple, Mary (2001). *Lexical Functional Grammar*. Bd. 34. Syntax and Semantic. New York: Academic Press.

Literatur II

- Declerck, Thierry (Mai 2006). „SynAF: Towards a Standard for Syntactic Annotation“. In: *Proceedings of the Fifth International Language Resources and Evaluation (LREC 2006)*. European Language Resources Association (ELRA). Genua, S. 229–232.
- DeRose, Steven J., David G. Durand, Elli Mylonas und Allen H. Renear (1990). „What is text, really?“ In: *Journal of Computing in Higher Education* 1.2, S. 3–26.
- Diewald, Nils, Maik Stührenberg, Anna Garbar und Daniela Goecke (2008). „Serengeti – Webbasierte Annotation semantischer Relationen“. In: *Journal for Language Technology and Computational Linguistics* 23.2, S. 74–93.
- Ide, Nancy M., Patrice Bonhomme und Laurent Romary (Mai 2000). „XCES: An XML-based Encoding Standard for Linguistic Corpora“. In: *Proceedings of the Second International Language Resources and Evaluation (LREC 2000)*. European Language Resources Association (ELRA). Athen, S. 825–830.
- ISO/TC 37/SC 3 (2009). *Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources*. International Standard ISO 12620:2009. Genf: International Organization for Standardization.
- ISO/TC 37/SC 4/WG 1 (2006). *Language Resource Management — Feature Structures – Part 1: Feature Structure Representation*. International Standard ISO 24610-1:2006. Genf: International Organization for Standardization.

Literatur III

ISO/TC 37/SC 4/WG 1 (Sep. 2011a). *Language Resource Management — Feature Structures – Part 2: Feature System Declaration*. International Standard ISO 24610-2:2011. Genf: International Organization for Standardization.

— (Aug. 2011b). *Language Resource Management — Linguistic annotation framework (LAF)*. Final Draft International Standard ISO/FDIS 24612. Genf: International Organization for Standardization.

ISO/TC 37/SC 4/WG 2 (2008). *Language Resource Management — Morpho-syntactic annotation framework*. Draft International Standard ISO/DIS 24611. Genf: International Organization for Standardization.

— (2010a). *Language Resource Management — Syntactic annotation framework (SynAF)*. International Standard ISO 24615:2010. Genf: International Organization for Standardization.

— (2010b). *Language Resource Management — Syntactic annotation framework (SynAF)*. Final Draft International Standard ISO/FDIS 24615. Genf: International Organization for Standardization. url: `http://www.tc37sc4.org/new_doc/iso_tc37_sc4_N712_wg2_FDIS_24615_SynAF_June2010-update.pdf` (besucht am 05.03.2012).

Kaplan, Ronald M. und Joan Bresnan (1982). „Lexical-Functional Grammar: A Formal System for Grammatical Representation“. In: *The Mental Representation of Grammatical Relations*. Hrsg. von Joan Bresnan. Cambridge: MIT Press, S. 173–281.

Marinelli, Paolo, Fabio Vitali und Stefano Zacchiroli (2008). „Towards the unification of formats for overlapping markup“. In: *New Review of Hypermedia and Multimedia* 14.1, S. 57–94.

Literatur IV

Mathet, Yann und Antoine Widlöcher (Juni 2011). „Stratégie d'exploration de corpus multi-annotés avec GlozzQL“. In: *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*. Hrsg. von Mathieu Lafourcade und Violaine Prince. Bd. 2. Association pour le Traitement Automatique des langues (ATALA). Montpellier, S. 143–148.

Pollard, Carl und Ivan A. Sag (1987). *Information-based Syntax and Semantics*. Menlo Park: CSLI, Center for the Study of Language and Information.

— (1994). *Head-Driven Phrase Structure Grammar*. Chicago: The University of Chicago Press.

Sperberg-McQueen, C. M. und Lou Burnard, Hrsg. (Nov. 1993). *TEI P2: Guidelines for the Encoding and Interchange of Machine Readable Texts*. Oxford u. a.: published for the TEI Consortium by Humanities Computing Unit, University of Oxford.

Sperberg-McQueen, C. M. und Claus Huitfeldt (1999). „Concurrent Document Hierarchies in MECS and SGML“. In: *Literary and Linguistic Computing* 14.1, S. 29–42.

— (2004). „GODDAG: A Data Structure for Overlapping Hierarchies“. In: *Digital Documents: Systems and Principles, 8th International Conference on Digital Documents and Electronic Publishing, DDEP 2000, 5th International Workshop on the Principles of Digital Document Processing, PODDP 2000, Munich, Germany, September 13-15, 2000, Revised Papers*. Hrsg. von Peter King und Ethan V. Munson. Bd. 2023. Lecture Notes in Computer Science 2023. Springer, S. 139–160.

Literatur V

Stührenberg, Maik, Daniela Goecke, Nils Diewald, Irene Cramer und Alexander Mehler (Juni 2007). „Web-based Annotation of Anaphoric Relations and Lexical Chains“. In: *Proceedings of the Linguistic Annotation Workshop*. Hrsg. von Branimir Boguraev, Nancy M. Ide, Adam Meyers, Shigeko Nariyama, Manfred Stede, Janyce Wiebe und Graham Wilcock. Prag: Association for Computational Linguistics, S. 140–147.

Stührenberg, Maik und Daniel Jettka (Aug. 2009). „A toolkit for multi-dimensional markup: The development of SGF to XStandoff“. In: *Proceedings of Balisage: The Markup Conference*. Bd. 3. Balisage Series on Markup Technologies. Montréal. doi: 10.4242/BalisageVol3.Stuhrenberg01.

Widlöcher, Antoine und Yann Mathet (Juni 2009). „La plate-forme Glozz : environnement d’annotation et d’exploration de corpus“. In: *Actes de la 16e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009) – Session posters*. Hrsg. von Mathieu Lafourcade und Violaine Prince. Association pour le Traitement Automatique des langues (ATALA). Senlis.

Witt, Andreas (Aug. 2004). „Multiple Hierarchies: New Aspects of an Old Solution“. In: *Proceedings of Extreme Markup Languages*. Montréal. url: http://conferences.idealliance.org/extreme/html/2004/Witt01/EML2004Witt01.html (besucht am 05. 03. 2012).