

Annotation multimodaler Dokumente

Maik Stührenberg

Institut für Deutsche Sprache
Universität Mannheim

Kolloquium Sprachwissenschaft

Was Sie erwartet:

- 1 Grundlagen linguistischer Annotation
 - Annotationsgrundlagen mit XML
 - Multiple Annotationen
- 2 Multimodale Dokumente
- 3 Multimodale Annotation
- 4 XStandoff
 - Segmentierung
 - Annotationen
- 5 Zusammenfassung

- 1 Grundlagen linguistischer Annotation
 - Annotationsgrundlagen mit XML
 - Multiple Annotationen
- 2 Multimodale Dokumente
- 3 Multimodale Annotation
- 4 XStandoff
- 5 Zusammenfassung

Das Konzept der Annotation

Annotation im Kontext

- In der Korpuslinguistik, Computerlinguistik und besonders in der Texttechnologie hängt die Verarbeitung von Sprachressourcen stark von der Annotation der Inhalte ab
- Annotation bedeutet das Hinzufügen (Auszeichnen) von relevanten Informationen/Markierungen zur den auszuzeichnenden Daten (den Primärdaten)

Ein Beispiel

Hey Paul! Would you give me
the hammer?

Annotation mit Hilfe der Metasprache XML

POS Annotation (Stanford NLP Tagger, XML-Ausgabe, angepasst)

```
<pos>
  <sentence id="s0">
    <word wid="s0w0" pos="NNP">Hey</word>
    <word wid="s0w1" pos="NNP">Paul</word>
    <word wid="s0w2" pos=".">!</word>
  </sentence>
  <sentence id="s1">
    <word wid="s1w0" pos="MD">Would</word>
    <word wid="s1w1" pos="PRP">you</word>
    <word wid="s1w2" pos="VB">give</word>
    <word wid="s1w3" pos="PRP">me</word>
    <word wid="s1w4" pos="DT">the</word>
    <word wid="s1w5" pos="NN">hammer</word>
    <word wid="s1w6" pos=".">?</word>
  </sentence>
</pos>
```

XML als Basis linguistischer Annotation

Als Metasprache stellt XML folgende Komponenten für Auszeichnungssprachen bereit (*Tripod-Modell* nach Sperberg-McQueen und Huitfeldt 1999; Huitfeldt und Sperberg-McQueen 2004)¹:

- Syntax zur Trennung von Auszeichnung und Primärdatum
- Formales Modell
- Formalismen zur Erstellung von Dokumentgrammatiken

Zur verbesserten Austauschbarkeit werden Auszeichnungssprachen auf Basis von Metasprachen entwickelt.

¹Zur Erweiterung des Tripod-Modells vgl. Stührenberg (2013a)

Grundlegende Einheiten der Informationsstrukturierung:

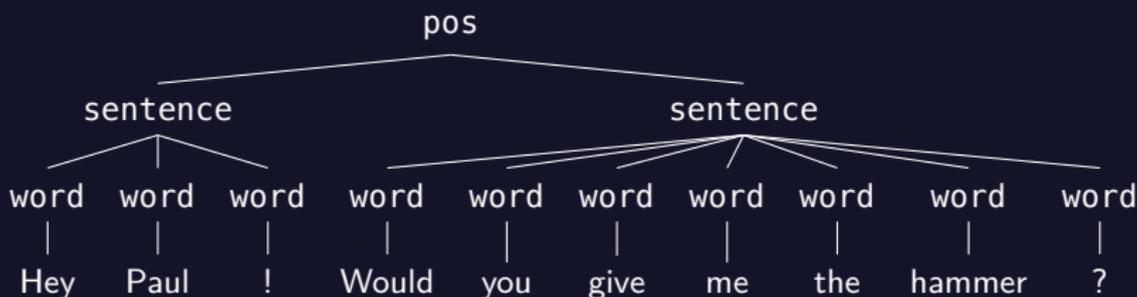
- Es gibt Elemente und Attribute
- Elemente bestehen aus Start- und Endtag, die den auszuzeichnenden Text (oder andere Elemente) umschließen
- Ein Starttag besteht aus der Zeichenfolge <Elementname>, ein Endtag aus der Zeichenfolge </Elementname>
- Elemente können andere Elemente (Containerelemente), Text (Datenelemente) oder beides (gemischtes Inhaltsmodell) beinhalten
- Attribute sind den Elementen zugeordnet, stehen im Starttag und folgen der Syntax `Attributname="Attributwert"`

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:
 - Es gibt genau ein Element, das alle anderen Elemente beinhaltet (Wurzelement)
 - 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 NLP-Taggerausgabe (ohne Attribute)



Anmerkung: Unter Verwendung von Link-Kanten in XML (per ID/IDREF) können auch Graphen-Strukturen abgebildet werden, hierzu ist aber eine Dokumentgrammatik notwendig (vgl. Stührenberg 2012)

XML: Dokumentgrammatik

Eine Dokumentgrammatik legt folgende Eigenschaften einer Auszeichnungssprache fest:

- Anzahl und Art der Elemente, d. h. Name, Inhaltsmodell, Datentyp, etc.
- Verschachtelung der Elemente untereinander
- Art und Anzahl der Attribute
- Zuordnung der Attribute zu Elementen

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

Vorteile XML-basierter Annotation

- Erstellung (nahezu) beliebiger oder Verwendung/Adaption vorhandener Auszeichnungssprachen
- Offener W3C-Standard
- XML-Instanzen sind Textdateien und lassen sich mit nahezu jedem Texteditor bearbeiten
- XML unterstützt Unicode (relevant für die linguistische Annotation)
- Standards wie XSLT (zur Transformation), XPath (zur Traversierung des Baummodells) und XQuery (Abfragesprache) sowie eine große Anzahl freier (und/oder Open Source) Software erleichtern den XML-Einsatz

Tipp: Viele von Ihnen nutzen XML bereits (DOCX, PPTX, XLSX,...)

Nachteile XML-basierter Annotation

- Steigerung der Dateigröße durch explizites Start- und Endtag (im Gegensatz z. B. zu JSON, vgl. Crockford 2006; ECMA International 2013)
- Darstellung multipler Annotationen problematisch
- Begrenzter semantischer Gehalt, d. h. einzig die Namen der Elemente und Attribute geben Hinweis auf die Motivation zur Auszeichnung – sofern nicht externe Taxonomien wie z. B. ISOcat² oder Ontologien zum Einsatz kommen

²<http://isocat.org>

Besonderheiten linguistischer Annotation

Das bekannte Beispiel

Hey Paul! Would you give me
the hammer?

Neben der bereits gezeigten POS-Annotation sind noch weitere Informationen von Bedeutung, z. B. die der Sprecher:

Sprecherinformationen

Peter: "Hey Paul! Would you give me"

Paul: "the hammer?"

Sollen die Äußerungen der Sprecher ebenfalls annotiert werden, überlappen diese mit der Satzstruktur.

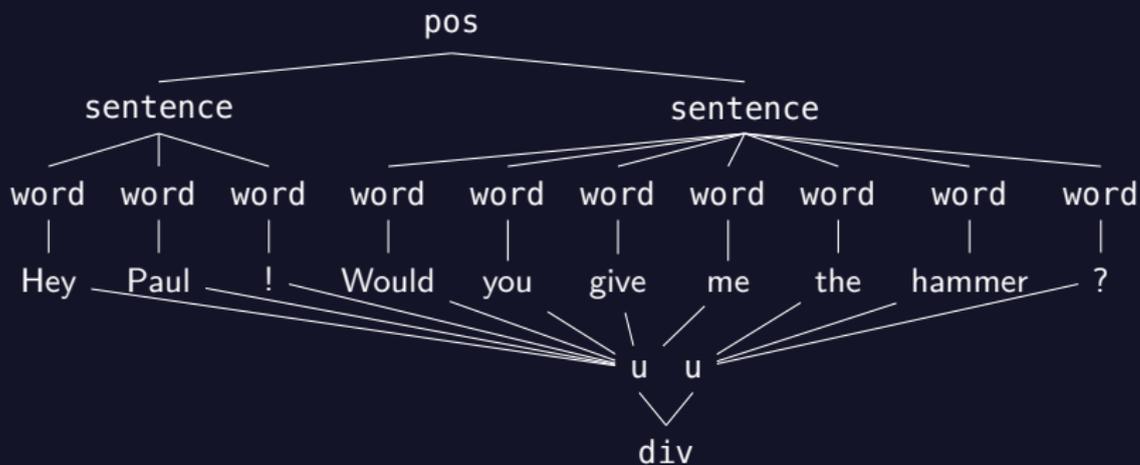
Multiple Annotationen

Kombination beider Annotationsebenen

```
<div type="dialog">
<pos>
<sentence id="s0">
  <u who="Peter">
    <word wid="s0w0" pos="NNP">Hey</word>
    <word wid="s0w1" pos="NNP">Paul</word>
    <word wid="s0w2" pos=".">!</word>
  </u>
</sentence>
<sentence id="s1">
  <word wid="w0" pos="MD">Would</word>
  <word wid="s1w1" pos="PRP">you</word>
  <word wid="s1w2" pos="VB">give</word>
  <word wid="s1w3" pos="PRP">me</word>
  </u>
  <u who="Paul">
    <word wid="s1w4" pos="DT">the</word>
    <word wid="s1w5" pos="NN">hammer</word>
    <word wid="s1w6" pos=".">?</word>
  </u>
</sentence>
</pos>
</div>
```

Multiple Annotationen

Beide Annotationsebenen in der grafischen Darstellung



Zwischenresultat

- Bei der linguistischen Annotation benötigt man oft mehrere Annotationsebenen
- Multiple Annotationen können *überlappen*, d. h. Elemente können multiple Elternknoten haben (Sperberg-McQueen und Huitfeldt 2004, S. 151)
- Überlappungen verstoßen gegen XMLs Wohlgeformtheitsbedingungen
- Verschiedene Lösungsansätze zur Darstellung überlappender Annotationen bzw. diskontinuierlicher Annotationseinheiten werden diskutiert:
 - Multiple Documente
 - Meilensteine
 - Fragmente
 - Standoff-Annotation

Multiple Dokumente

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

Dokument 1

```
<pos>
<sentence id="s0">
  <word wid="s0w0" pos="NNP">Hey</word>
  <word wid="s0w1" pos="NNP">Paul</word>
  <word wid="s0w2" pos=".">!</word>
</sentence>
<sentence id="s1">
  <word wid="s1w0" pos="MD">Would</word>
  <word wid="s1w1" pos="PRP">you</word>
  <word wid="s1w2" pos="VB">give</word>
  <word wid="s1w3" pos="PRP">me</word>
  <word wid="s1w4" pos="DT">the</word>
  <word wid="s1w5" pos="NN">hammer</word>
  <word wid="s1w6" pos=".">?</word>
</sentence>
</pos>
```

Dokument 2

```
<div type="dialog">
  <u who="Peter">Hey Paul! Would you give me</u>
  <u who="Paul">the hammer?</u>
</div>
```

Multiple Dokumente

Vorteile

- Jede Datei ist vollständig und einzeln verwendbar
- Sowohl für menschliche Leser als auch Software gut nachvollziehbar
- 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: Markieren des Beginns (und implizit des Endes) einer Annotation mit Hilfe spezieller leerer Elemente (definiert erstmals in TEI P2, Abschnitt 22.3.4.3, „Milestone method“)

Meilensteine nach TEI P5

```
<pos>
<sentence id="s0">
  <milestone unit="u" n="Peter"/>
  <word wid="s0w0" pos="NNP">Hey</word>
  <word wid="s0w1" pos="NNP">Paul</word>
  <word wid="s0w2" pos=".">!</word>
</sentence>
<sentence id="s1">
  <word wid="s1w0" pos="MD">Would</word>
  <word wid="s1w1" pos="PRP">you</word>
  <word wid="s1w2" pos="VB">give</word>
  <word wid="s1w3" pos="PRP">me</word>
  <milestone unit="u" n="Paul"/>
  <word wid="s1w4" pos="DT">the</word>
  <word wid="s1w5" pos="NN">hammer</word>
  <word wid="s1w6" pos=".">?</word>
</sentence>
</pos>
```

Meilensteine

Vorteile

- Ursprüngliche Elementinformationen und -typen können in Form von Attributen gespeichert werden
- Unterscheidung zwischen generischen Elementen und solchen mit eigenem semantischen Gehalt (1b) möglich

Nachteile

- Bei vielfacher Überlappung unübersichtlich
- Besondere Formen von Überlappung (z. B. *Self Overlap*) nicht abbildbar (für eine Typologie vgl. Witt u. a. 2005)
- Schlecht maschinell zu verarbeiten
- Keine Unterscheidung von Inklusion und Dominanz (vgl. Sperberg-McQueen und Huitfeldt 2008; Stührenberg und Jettka 2009)

Fragmentierungen

Vorgehensweise: Aufbrechen überlappender Elemente in kleinere Teilfragmente (vgl. P5 2.7.0, Kapitel 20.3)

Fragmentierung nach TEI P5, Variante A

```
<pos>
<u who="Peter">
<sentence id="s0" n="sentence1">
  <word wid="s0w0" pos="NNP">Hey</word>
  <word wid="s0w1" pos="NNP">Paul</word>
  <word wid="s0w2" pos=".">!</word>
</sentence>
<sentence id="s1" n="sentence2">
  <word wid="s1w0" pos="MD">Would</word>
  <word wid="s1w1" pos="PRP">you</word>
  <word wid="s1w2" pos="VB">give</word>
  <word wid="s1w3" pos="PRP">me</word>
</sentence>
</u>
<u who="Paul">
  <sentence id="s2" n="sentence2">
    <word wid="s1w4" pos="DT">the</word>
    <word wid="s1w5" pos="NN">hammer</word>
    <word wid="s1w6" pos=".">?</word>
  </sentence>
</u>
</pos>
```

Fragmentierungen

Vorteile

- Durch den Wert des n-Attributs wird die logische Zusammengehörigkeit der Fragmente deutlich gemacht

Nachteile

- Zusätzliche Elemente werden eingefügt
- Teilweise semantisch irreführend, dadurch schlecht maschinell zu verarbeiten
- Keine Unterscheidung von Inklusion und Dominanz

Fragmentierungen

Fragmentierung nach TEI P5, Variante B (Chaining)

```
<pos>
<u who="Peter">
<sentence id="s0">
  <word wid="s0w0" pos="NNP">Hey</word>
  <word wid="s0w1" pos="NNP">Paul</word>
  <word wid="s0w2" pos=".">!</word>
</sentence>
<sentence id="s1" next="#s2">
  <word wid="s1w0" pos="MD">Would</word>
  <word wid="s1w1" pos="PRP">you</word>
  <word wid="s1w2" pos="VB">give</word>
  <word wid="s1w3" pos="PRP">me</word>
</sentence>
</u>
<u who="Paul">
  <sentence id="s2" prev="#s1">
    <word wid="s1w4" pos="DT">the</word>
    <word wid="s1w5" pos="NN">hammer</word>
    <word wid="s1w6" pos=".">?</word>
  </sentence>
</u>
</pos>
```

Fragmentierungen

Vorteile

- Durch Hinzufügen der `next`- und `prev`-Attribute wird die logische Zusammengehörigkeit der Fragmente deutlich gemacht

Nachteile

- Bei vielfacher Überlappung unübersichtlich
- Schlecht maschinell zu verarbeiten
- Keine Unterscheidung von Inklusion und Dominanz

Fragmentierungen

Fragmentierung nach TEI P5, Variante C (Virtual Join)

```
<pos>
  <u who="Peter">
    <word wid="s0w0" pos="NNP">Hey</word>
    <word wid="s0w1" pos="NNP">Paul</word>
    <word wid="s0w2" pos=".">!</word>
    <word wid="s1w0" pos="MD">Would</word>
    <word wid="s1w1" pos="PRP">you</word>
    <word wid="s1w2" pos="VB">give</word>
    <word wid="s1w3" pos="PRP">me</word>
  </u>
  <u who="Paul">
    <word wid="s1w4" pos="DT">the</word>
    <word wid="s1w5" pos="NN">hammer</word>
    <word wid="s1w6" pos=".">?</word>
  </u>
  <!-- ... -->
  <join results="sentence" scope="root" target="#s0w0 #s0w1 #s0w2"/>
  <join results="sentence" scope="root" target="#s1w0 #s1w1 #s1w2 #s1w3 #s1w4 #s1w5 #s1w6"/>
</pos>
```

Fragmentierungen

Vorteile

- Explizite und implizite Hierarchien können abgebildet werden

Nachteile

- Eine Hierarchieebene wird bevorzugt
- Virtuelles Zusammenfügen (*virtual join*) der impliziten Hierarchie erfordert eine entsprechende Verarbeitung

Standoff-Notation

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

Mögliche (nicht zwingend ideale) Realisierung

```
<root>
  <div 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="s0" start="0" end="9"/>
  <sentence id="s1" start="10" end="39"/>
  <w wid="s0w0" pos="NNP" start="0" end="3"/>
  <w wid="s0w1" pos="NNP" start="4" end="8"/>
  <w wid="s0w2" pos="." start="8" end="9"/>
  <w wid="s1w0" pos="MD" start="10" end="15"/>
  <w wid="s1w1" pos="PRP" start="16" end="19"/>
  <w wid="s1w2" pos="VB" start="20" end="24"/>
  <w wid="s1w3" pos="PRP" start="25" end="27"/>
  <w wid="s1w4" pos="DT" start="28" end="31"/>
  <w wid="s1w5" pos="NN" start="32" end="38"/>
  <w wid="s1w6" pos="." start="38" end="39"/>
</root>
```

Standoff-Notation

Vorteile

- Nelson (1997) nennt eine Reihe von Gründen (u. a. auch rechtlicher Art)
- Beliebig viele Annotationsebenen kombinierbar
- Prinzipiell gut skalierbar
- Verschiedene Serialisierungsformate vorstellbar – inkl. Abbildung von Graphen und Unterscheidung von Inklusion und Dominanzbeziehung (vgl. Stührenberg und Jettka 2009)

Nachteile

- Für menschliche Leser sehr schlecht nachvollziehbar
- Maschinelle Verarbeitung nicht trivial
- Je nach Ansatz geringe Robustheit bzgl. Primärdatenintegrität

Standoff-Annotation – ein neuer Ansatz?

Geschichtliche Entwicklung

- Standoff-Annotation wurde im Rahmen des TIPSTER-Projekts entwickelt (TIPSTER Text Phase II Architecture Concept)
- Der Corpus Encoding Standard (CES, Ide u. a. 1996; Ide 1998) ermöglichte die TEI-P3-kompatible Standoff-Annotation
- Standoff findet sich in der Text Encoding Initiative (TEI) unter Verwendung von XInclude (Marsh u. a. 2006) und dem XPointer element() scheme (Grosso u. a. 2003)
- Das TEI Feature Structure Format (P5 2.7.0; Stegmann und Witt 2009), standardisiert als ISO 24610-1:2006, setzt ebenfalls auf Standoff

1 Grundlagen linguistischer Annotation

2 Multimodale Dokumente

3 Multimodale Annotation

4 XStandoff

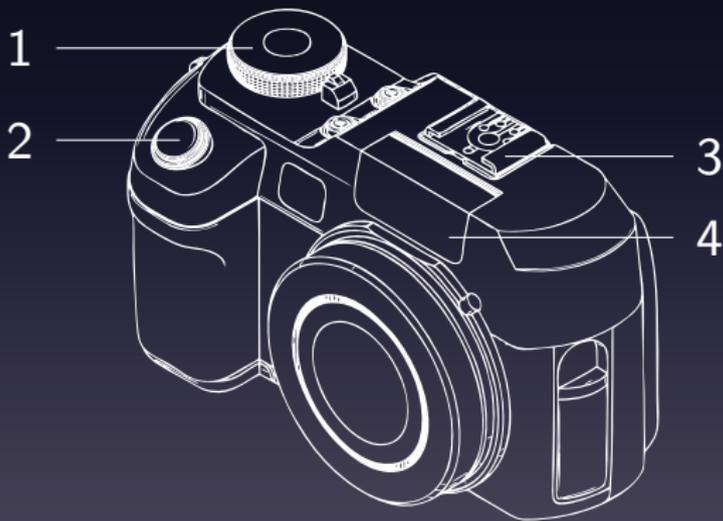
5 Zusammenfassung

Multimodale Dokumente – ein Einstieg

- Im Laufe dieses Vortrags nennen wir die digital vorliegenden Dokumente, die multiple Informationskodierungen enthalten, *multimodale Dokumente* (für detaillierte Begriffsdefinitionen vgl. Fricke 2012; Lücking und Pfeiffer 2012; Stöckl 2011)
- Als multiple Informationskodierungen bezeichnen wir eine Kombination aus selektierbarem Text (d. h. nicht als Bild gespeichert) und einer oder mehreren graphischen Darstellungen (oder anderer nicht-textueller Kodierungen)
- Multiple Kodierungen (Repräsentationsformen) sind oftmals nicht nur alternative Kodierungen der selben Informationen
- Verschiedene Kodierungen können in Beziehungen stehen
- Beziehungen können interessant sein

Fangen wir mit einem einfachen Beispiel an...

- 1 Stellen sie das Modusrad (1) auf die Stellung 'A' oder 'P'
- 2 Drücken sie den Auslöser (2) halb herunter, um zu Fokussieren und den internen Blitz (4) auszufahren
Alternativ verwenden sie einen externen Blitz mit Hilfe des Blitzschuhs (3)
- 3 Drücken sie den Auslöser (2) ganz herunter, um das Bild aufzunehmen

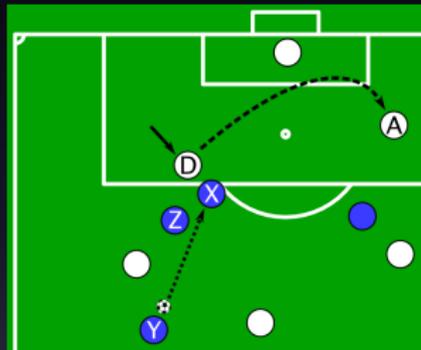


Gebrauchsanweisung

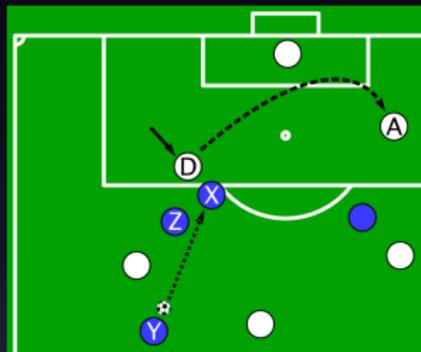
- Typische Anweisung für das Fotografieren mit einer digitalen Kamera enthalten sowohl eine grafische Darstellung der Kamera als auch textuelle Beschreibungen
- Die Textanweisungen beziehen sich auf Teile der Grafik (durch Nummern)
- Teile der Grafik (Regionen) bezeichnen die entsprechenden Bedienelemente der Kamera in der realen Welt
- Wollten wir solche multimodalen Texte annotieren (und in einem zweiten Schritt verarbeiten), bräuchten wir ein entsprechendes universelles Format.

Ein etwas interessanteres Beispiel...

Eine weitere Situation, in der Team A es nicht schafft, seinen Zug zu beenden: Y versucht den Ball durch die schmale Lücke zu X zu passen (anstatt ihn zu G zu spielen), während Z unfreiwillig Y blockiert. Aber bevor der Ball X erreicht, fängt ihn D ab und passt den Ball zu A.



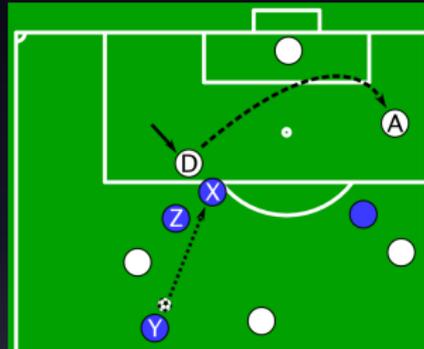
Eine weitere Situation, in der **Team A** es nicht schafft, seinen Zug zu beenden: **Y** versucht den Ball durch **die schmale Lücke** zu **X** zu passen (anstatt ihn zu **G** zu spielen), während **Z** unfreiwillig **Y** blockiert. Aber bevor der Ball **X**, erreicht, fängt ihn **D** ab und passt den Ball zu **A**.



Text und Bild beinhalten verschiedene Informationen...

- ...über **Teams**
- ...**Spieler**
- ...**Orte**

Eine weitere Situation, in der Team A es nicht schafft, seinen Zug zu beenden: Y versucht den Ball durch **die schmale Lücke** zu X zu passen (anstatt ihn zu G zu spielen), während Z unfreiwillig Y blockiert. Aber bevor der Ball X erreicht, fängt ihn D ab und passt den Ball zu A.



Nicht jede Informationseinheit ist in beiden Kodierungen vorhanden:

- Wo ist der Spieler „G“?
- Welcher Ort ist die „schmale Lücke“?

1 Grundlagen linguistischer Annotation

2 Multimodale Dokumente

3 Multimodale Annotation

4 XStandoff

5 Zusammenfassung

Annotation der Fußballanalyse

- Zur Annotation multimodaler Dokumente müssen wir zunächst Text und Bild trennen und separat annotieren
- Eine einfache Inline-Annotation des Textes kann z. B. die folgenden Elemente und Attribute verwenden:

Textuelle Primärdaten

```
<text>
```

Eine weitere Situation, in der `<team name="teamA">Team A</team>` es nicht schafft, seinen Zug zu beenden: `<player name="Y" team="teamA">Y</player>` versucht den Ball durch `<place name="gap">die schmale Lücke</place>` zu `<player name="X" team="teamA">X</player>` zu passen (anstatt ihn zu `<player name="G" team="teamA">G</player>` zu spielen), während `<player name="Z" team="teamA">Z</player>` unfreiwillig `<player name="Y" team="teamA">Y</player>` blockiert.

Aber bevor der Ball `<player name="X" team="teamA">X</player>` erreicht, fängt ihn `<player name="D" team="teamB">D</player>` ab und passt den Ball zu `<player name="A" team="teamB">A</player>`.

```
</text>
```

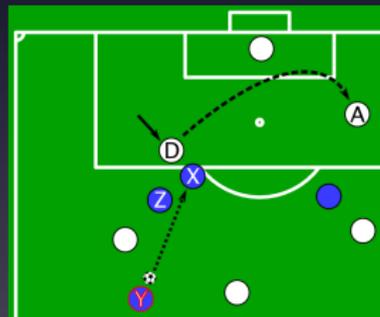
Annotation der Grafik

- Es gibt eine Reihe von Spezifikationen zur Annotation von Grafiken oder auch Grafiken und Text (Henschel 2003) – diese behandeln Grafiken aber als atomare Einheiten
- Es gibt Arbeiten zur Syntax von Grafiken und deren logische Bestandteilen wie z. B. von Engelhardt (2002)
- Wir benötigen ein System, um *Teile* einer Grafik zu adressieren, wenn wir diese Teilgrafiken als Gegenstand eines Diskurses verarbeiten wollen

Annotation der Grafik

Zur Annotation der Grafik gehen wir von folgenden Annahmen aus:

- Die vollständige Grafik wird über eine URL (lokal/fern) erreicht
- Einzelne Punkte lassen sich über ein Koordinatensystem adressieren
- In Kombination mit grundlegenden Formen ermöglichen Punkte die Auswahl einzelner Bildteile
- Ein Kreis wird beschrieben durch das Koordinatenpaar x_{center} , y_{center} und den Radius r
- Der den Spieler Y repräsentierende Kreis kann durch die Angaben $x_{center}=138$, $y_{center}=278$, und $r=9$ selektiert werden (Angaben in px, 0,0-Koordinate oben links)



Wie serialisieren?

HTML *image maps* (Verweis-sensitive Grafiken) können zur Serialisierung herangezogen werden:

- 1 Das `img` Element referenziert die Bilddatei
- 2 Das `map` Element enthält die Definition einer *image map*
- 3 Das `area` Element definiert einzelne Bildteile als Verweis-sensitiv und enthält das Linkziel

```

<map name="map">
  <area shape="circle" coords="138,278,9" href="#playerY" alt="Player Y" title="Player Y"/>
</map>
```

Weitere mögliche Serialisierungen sind TEI's `zone` Element (P5 2.7.0, 11.4, „Advanced uses of Surface and Zone“) oder SVG's Grundformen und Pfade (Dahlström u. a. 2011)

Zwischenstand

Was haben wir bis jetzt?

- Wir können Text und Bild separat annotieren
- Text wird inline annotiert, Bilder müssen generell standoff annotiert werden

Wo wollen wir hin?

Wünschenswert wäre ein integratives Format mit Unterstützung für...

- multimodale Dokumente und
- multiple Annotationsebenen

1 Grundlagen linguistischer Annotation

2 Multimodale Dokumente

3 Multimodale Annotation

4 XStandoff

- Segmentierung
- Annotationen

5 Zusammenfassung

- Ursprünglich entwickelt als Sekimo Generic Format (SGF) im Rahmen des Sekimo-Projekts (Stührenberg und Goecke 2008)
- Standoff *Meta*-Annotationsformat
- Unterstützt multiple Primärdaten
- Segmentierungsmechanismen für verschiedene Primärdatentypen
- Unterscheidung zwischen Level (konzeptuelle Ebene) und Layer (Serialisierung)
- Multiple Annotationsebenen – keine Einschränkungen bzgl. Markupinventar (Elemente/Attribute)
- ISOcat-Attribute für importierte Annotationsbenen
- XSD 1.1 Schema (seit XStandoff 2.1)
- XStandoff Toolkit (Stührenberg und Jettka 2009)
- Open Source (GNU Lesser General Public License, LGPL v3)

XStandoff-Struktur (ohne Metadaten)

```
<corpusData xml:id="c1">
  <primaryData xml:id="p1">
    <!-- Primärdatenreferenz (mehrfache Vorkommen möglich) -->
  </primaryData>
  <segmentation>
    <!-- segments -->
  </segmentation>
  <annotation>
    <level>
      <layer>
        <!-- Annotationsebene(n) (Serialisierung) -->
      </layer>
    </level>
    <!-- mögliche weitere Annotationsebene(n) (konzeptuelle Ebene) -->
  </annotation>
  <!-- mögliche weitere corpusData-Elemente -->
</corpusData>
```

XStandoff-Struktur (ohne Metadaten)

```
<corpusData xml:id="c1">
  <primaryData xml:id="p1">
    <!-- Primärdatenreferenz (mehrfache Vorkommen möglich) -->
  </primaryData>
  <segmentation>
    <!-- segments -->
  </segmentation>
  <annotation>
    <level>
      <layer>
        <!-- Annotationsebene(n) (Serialisierung) -->
      </layer>
    </level>
    <!-- mögliche weitere Annotationsebene(n) (konzeptuelle Ebene) -->
  </annotation>
  <!-- mögliche weitere corpusData-Elemente -->
</corpusData>
```

XStandoff-Struktur (ohne Metadaten)

```
<corpusData xml:id="c1">
  <primaryData xml:id="p1">
    <!-- Primärdatenreferenz (mehrfache Vorkommen möglich) -->
  </primaryData>
  <segmentation>
    <!-- segments -->
  </segmentation>
  <annotation>
    <level>
      <layer>
        <!-- Annotationsebene(n) (Serialisierung) -->
      </layer>
    </level>
    <!-- mögliche weitere Annotationsebene(n) (konzeptuelle Ebene) -->
  </annotation>
  <!-- mögliche weitere corpusData-Elemente -->
</corpusData>
```

XStandoff-Struktur (ohne Metadaten)

```
<corpusData xml:id="c1">
  <primaryData xml:id="p1">
    <!-- Primärdatenreferenz (mehrfache Vorkommen möglich) -->
  </primaryData>
  <segmentation>
    <!-- segments -->
  </segmentation>
  <annotation>
    <level>
      <layer>
        <!-- Annotationsebene(n) (Serialisierung) -->
      </layer>
    </level>
    <!-- mögliche weitere Annotationsebene(n) (konzeptuelle Ebene) -->
  </annotation>
  <!-- mögliche weitere corpusData-Elemente -->
</corpusData>
```

XStandoff-Struktur (ohne Metadaten)

```
<corpusData xml:id="c1">
  <primaryData xml:id="p1">
    <!-- Primärdatenreferenz (mehrfache Vorkommen möglich) -->
  </primaryData>
  <segmentation>
    <!-- segments -->
  </segmentation>
  <annotation>
    <level>
      <layer>
        <!-- Annotationsebene(n) (Serialisierung) -->
      </layer>
    </level>
    <!-- mögliche weitere Annotationsebene(n) (konzeptuelle Ebene) -->
  </annotation>
  <!-- mögliche weitere corpusData-Elemente -->
</corpusData>
```

Baustein 1: Segmentierung

Segmentierung in XStandoff ist abhängig vom Primärdatentyp – XStandoff 2.1 unterstützt:

- textuelle Primärdaten
- audio-visuelle Primärdaten
- grafische Primärdaten (vgl. Stührenberg 2013b)
- vor-annotierte Primärdaten (z. B. Webseiten, vgl. Stührenberg 2014)

Baustein 1: Segmentierung

Textuelle Primärdaten können mittels Zeichenpositionen unterteilt werden:

```
T h i s   i s   a   w o r d  
00|01|02|03|04|05|06|07|08|09|10|11|12|13|14
```

Serialisierung in XStandoff

```
<primaryData xml:id="txt">  
  <xsf:primaryDataRef uri="soccer.txt" encoding="utf-8" mimeType="text/plain" start="0"  
    end="265"/>  
</primaryData>  
<!-- [...] -->  
<segment xml:id="seg_text1" primaryData="txt" type="char" start="27" end="33"/>
```

Baustein 1: Segmentierung

Multimediale Primärdaten (Audio/Video) können über Zeitpunkte segmentiert werden:

Serialisierung in XStandoff

```
<primaryData start="0" end="335587" unit="milliseconds" xml:id="pd1_video">  
  <primaryDataRef uri="b1-video.mpg" mimeType="video/mpeg"/>  
</primaryData>  
<!-- [...] -->  
<segment xml:id="seg2" primaryData="pd1_video" start="309924" end="310079"/>
```

Baustein 1: Segmentierung

Grafiken lassen sich mit Hilfe eines Koordinatensystems und grundlegenden Formen unterteilen:

Serialisierung eines Kreises in XStandoff

```
<xsf:primaryData xml:id="img" unit="pixels">
  <xsf:primaryDataRef uri="img.png" mimeType="image/png" width="824" height="679"/>
</xsf:primaryData>
<!-- [...] -->
<xsf:segment xml:id="seg1" type="spatial" primaryData="img" shape="circle" coords="312,651,23"
  name="X"/>
```

Serialisierung eines Polygons in XStandoff

```
<xsf:segment xml:id="seg2" type="spatial" primaryData="img" shape="poly"
  coords="2400,125 2600,125 2400,945 2600,945" name="Y"/>
```

Baustein 1: Segmentierung

Komplexere grafische Formen lassen sich mittels Bézierkurven (hier: Adaption der SVG-Syntax) beschreiben

Serialisierung einer komplexen Form in XStandoff

```
<xsf:segment xml:id="seg2" type="spatial" shape="bezier" coords="C2400,125 2600,125 2400,945  
2600,945" primaryData="img" name="Z"/>
```

Baustein 1: Segmentierung

Die Kombination aus den Serialisierungsmechanismen für multimediale und grafische Primärdaten erlaubt die Annotation einzelner Bildteile (z. B. Körperteile in der Gestenannotation) über einen Zeitraum:

Serialisierung in XStandoff (Variante A)

```
<segment xml:id="s1" type="spatial" shape="poly" coords="0,11,20 3,4,30 1,2,30"
start="00:20:00" end="00:20:01"/>
```

Serialisierung in XStandoff durch Referenzierung bestehender Segmente

```
<segment xml:id="s1" type="spatial" shape="poly" coords="0,10,30 100,150,30 0,200,30"/>
<segment xml:id="s2" type="spatial" shape="poly" coords="10,10,40 110,150,40 10,200,40
110,100,40"/>
<segment xml:id="s3" type="seg" segments="s1 s2" name="AnkleLeft" mode="continuous"
start="00:00:00" end="00:01:15"/>
```

Baustein 1: Segmentierung

Obwohl die Einbeziehung bestehender Annotation in die Berechnung von Zeichenpositionen möglich ist, ist die Verwendung von XPath-2.0-Ausdrücken (Berglund u. a. 2010) zur Auswahl von Teilbäumen wohlgeformter XML-Instanzen sinnvoller:

XHTML-Primärdaten

```
<html xmlns="http://www.w3.org/1999/xhtml">
  <head><title>Instance</title></head>
  <body><div>This is a word.</div></body>
</html>
```

Serialisierung in XStandoff

```
<segment xml:id="seg_html1" primaryData="pd1" target="xhtml:html/xhtml:body/xhtml:div[1]"/>
```

Baustein 1: Segmentierung

Mit Hilfe von XQuery 1.0 und XPath 2.0 Funktionen (Malhotra u. a. 2010), lassen sich auch Teilzeichenketten eines Teilbaums auswählen:

XHTML-Primärdaten

```
<html xmlns="http://www.w3.org/1999/xhtml">
  <head><title>Instance</title></head>
  <body><div>This is a word.</div></body>
</html>
```

Serialization in XStandoff

```
<segment xml:id="seg2" primaryData="p1" target="xhtml:html/xhtml:body/substring(xhtml:div[1],1,4)"/>
```

Baustein 2: Annotationen

Während andere Standoff-Formate oftmals generische Auszeichnungsformate verwenden, bleibt XStandoff so nahe wie möglich an einer ursprünglichen/idealen Inline-Notation:

Inline

```
<text>
```

```
Eine weitere Situation, in der <team name="teamA">Team A  
</team> es nicht schafft, seinen Zug zu beenden: <player  
name="Y" team="teamA">Y</player> versucht den Ball  
durch <place name="gap">die schmale Lücke</place> zu  
<player name="X" team="teamA">X</player> zu passen  
(anstatt ihn zu <player name="G" team="teamA">G</  
player> zu spielen), während <player name="Z" team="  
teamA">Z</player> unfreiwillig <player name="Y" team  
="teamA">Y</player> blockiert.
```

```
<!-- [...] -->  
</text>
```

XStandoff

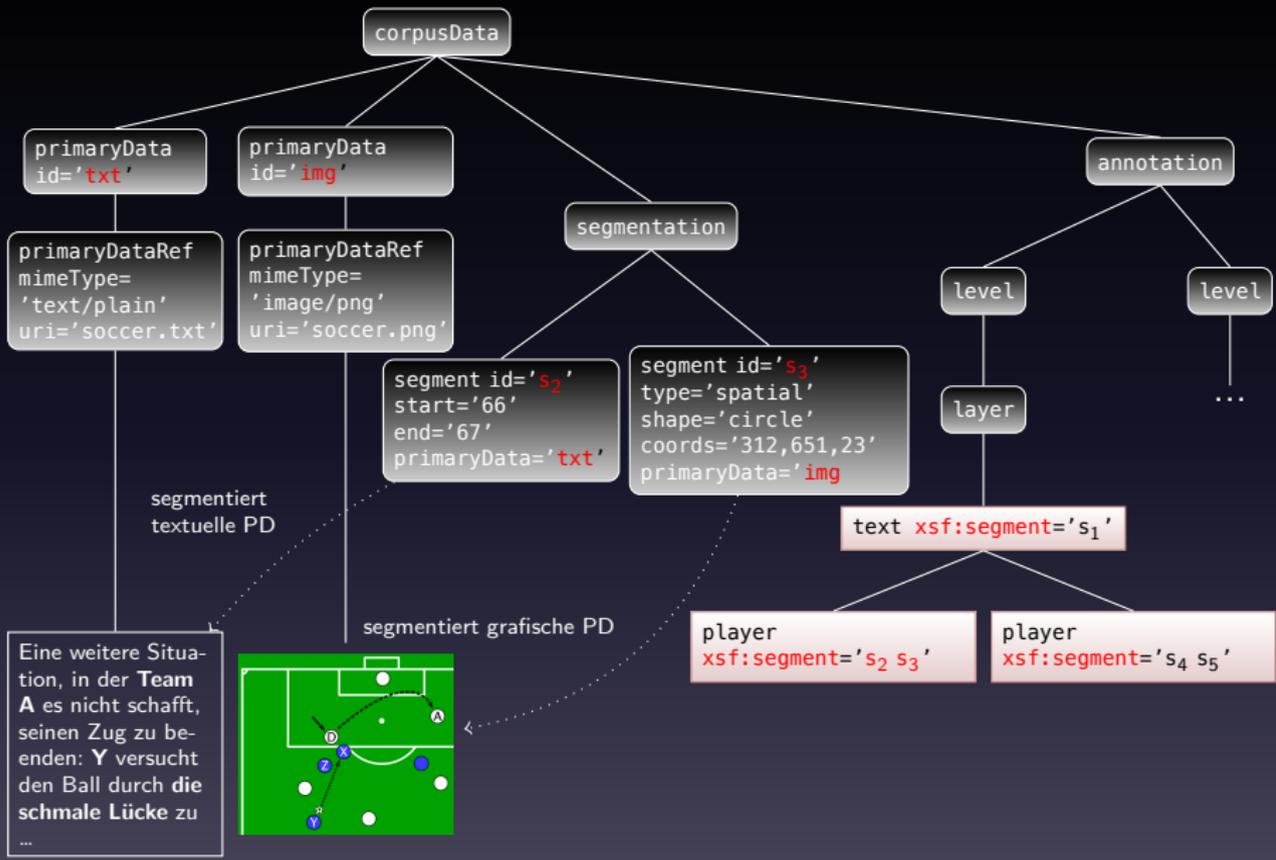
```
<text xsf:segment="seg1">  
  <team name="teamA" xsf:segment="seg2  
    seg19"/>  
  <player name="Y" team="teamA"  
    xsf:segment="seg3 seg12"/>  
  <place name="gap" xsf:segment="seg4  
    seg18"/>  
<!-- [...] -->  
</text>
```

Die Hierarchie der Elemente sowie alle Attribute bleiben erhalten, nur der textuelle Inhalt wird entfernt und XSF-eigene Attribute werden hinzugefügt

Kombination der Bausteine...

Verknüpfung von Segmenten und Annotationen per ID/IDREFS

```
<corpusData>
  <primaryData xml:id="txt">
    <primaryDataRef uri="soccer.txt" encoding="utf-8" mimeType="text/plain" start="0" end="265"/>
  </primaryData>
  <primaryData xml:id="img" unit="pixels">
    <primaryDataRef uri="soccer.png" mimeType="image/png" width="362" height="321"/>
  </primaryData>
  <segmentation>
    <!-- [...] -->
    <segment xml:id="seg3" start="66" end="67" primaryData="txt"/>
    <segment xml:id="seg12" type="spatial" shape="circle" coords="138,278,9" primaryData="img"
      name="Y"/>
    <segment xml:id="seg18" type="spatial" shape="poly" coords="142,246 167,185 188,194 163,255"
      primaryData="img" name="gap"/>
    <segment xml:id="seg19" type="seg" segments="seg12 seg13 seg14 seg17" mode="disjoint" name="
      Team A"/>
  </segmentation>
  <annotation>
    <level xml:id="soccer_a_b-level1">
      <layer>
        <text xsf:segment="seg1">
          <player name="teamA" xsf:segment="seg2 seg19"/>
          <player name="Y" team="teamA" xsf:segment="seg3 seg12"/>
        <!-- [...] -->
      </layer>
    </level>
  </annotation>
</corpusData>
```



Annotation der Relationen von Text und Bild

Einfach: Referenz beider Segmente und Hinzufügen von Metadaten

```
<player name="G" team="teamA" xsf:segment="seg6 seg17">
  <xsf:meta xmlns="http://www.tei-c.org/ns/1.0">
    <certainty locus="name" target="playerG" degree="0.9">
      <desc>Although the part of the graphic depicted with seg17 comes without a name tag, it is
        most likely that it depicts the player called 'G' in the running text (seg6).</desc>
    </certainty>
  </xsf:meta>
</player>
```

Ausführlich: Verwendung einer eigenen Annotationsebene

```
<!-- [...] -->
<level xml:id="soccer_relations">
  <layer>
    <relations>
      <isFormatOf xsf:segment="seg6 seg17" certainty="0.9"/>
      <!-- [...] -->
    </relations>
  </layer>
</level>
```

Annotation der Relationen von Text und Bild

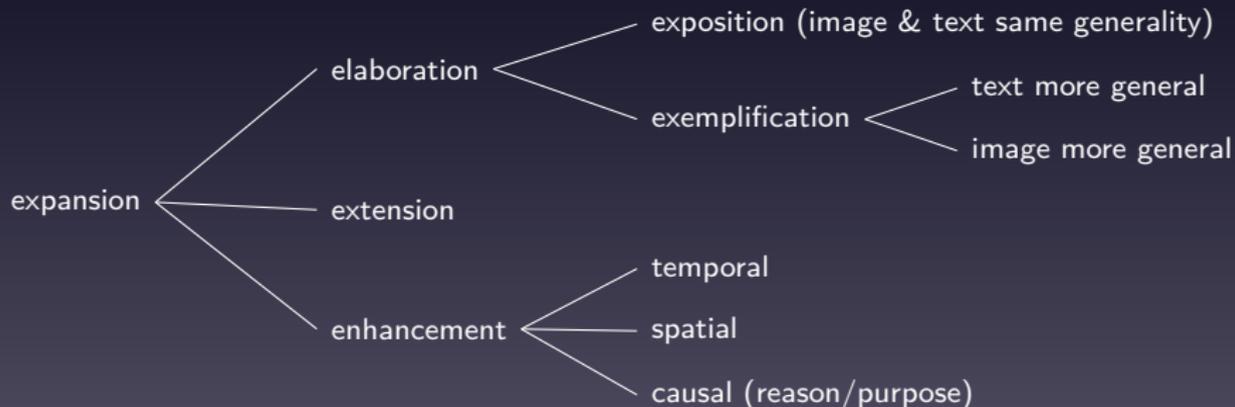
- XStandoff macht keine Einschränkungen bzgl. der Relationen und deren Serialisierung
- Kandidaten sind Serialisierungen der Text-Bild-Relationen aus Martinec und Salway (2005), van Leeuwen (2005) und Stöckl (2009), der RST-Relationen aus dem GeM-Projekt (Henschel 2003) oder ein Adaption der klassischen RST (Mann und Thompson 1988)
- Ein klassisches Forschungsfeld ist Text-Bild-Semiotik bzw. Bildlinguistik mit Anwendungen in der Werbung, Filmanalyse oder IBK, sowie die automatische Layouterkennung – für Fragestellungen vgl. Stöckl (2008)

Beispielrelationen aus Martinec und Salway (2005)

Bild-Text-Statusrelationen zwischen *gesamten* Bild und Text(-teilen)



Expansion-Relation



Exkurs: Annotation multimodaler Wikipedia-Artikel

Vorgehensweise bei der Annotation multimodaler Wikipedia-Artikel:

- 1 Wikipedia-Artikel liegen in Form von Wikitext-Markup vor (nicht XML-basiert und daher nicht wohlgeformt)
- 2 (X)HTML-Serialisierungen sind prinzipiell verwendbar, oftmals aber ebenfalls nicht wohlgeformt → Konvertierung nach WikiXML (vorzugsweise) oder I5 (Margareta und Längen 2014)
- 3 Konvertierung nach XStandoff mittels XStandoff-Toolkit
- 4 Erfassung der eingebetteten Bilder als XSF-Primärdaten
- 5 Erstellen einer zusätzlichen Annotationsebene zur Kodierung der Relationen

1 Grundlagen linguistischer Annotation

2 Multimodale Dokumente

3 Multimodale Annotation

4 XStandoff

5 Zusammenfassung

Zusammenfassung

- Multimodale Dokumente stellen besondere Anforderungen an eine XML-basierte Annotation dar, sind aber interessante Gegenstände verschiedener linguistischer Betrachtungen
- Standoff-Notation kann eine wertvolle Methode zur Annotation multimodaler Dokumente sein
- XStandoff kann als Metaformat für eine solche Annotation eingesetzt werden, da es eine Reihe an Segmentierungen und multiple Annotationsebenen unterstützt
- ☹️ Aktuell fehlt es noch an einem komfortablen Werkzeug zur Annotation multimodaler Dokumente, Vorarbeiten existieren aber bereits in Form von *Serengeti* (Stührenberg, Goecke u. a. 2007; Diewald u. a. 2008), der *Text-Image Linking Environment* (TILE)³ oder dem *Image Markup Tool* (IMT)⁴

³<http://mith.umd.edu/tile/>

⁴http://tapor.uvic.ca/~mholmes/image_markup/

Vielen Dank für die Aufmerksamkeit!

Fragen? Anmerkungen?

stuehrenberg@ids-mannheim.de | maik@xstandoff.net
<http://xstandoff.net>

Bibliographie I



Architecture Committee for the TIPSTER Text Phase II Program (1996). *TIPSTER Text Phase II Architecture Concept*. Techn. Ber. National Institute of Standards and Technology. URL:

http://www-nlpir.nist.gov/related_projects/tipster/docs/con112.doc
(besucht am 20.12.2011).



Berglund, Anders, Scott Boag, Don Chamberlin, Mary F. Fernández, Michael Kay, Jonathan Robie und Jérôme Siméon (2010). *XML Path Language (XPath). Version 2.0 (Second Edition)*. W3C Recommendation. World Wide Web Consortium (W3C). URL: <http://www.w3.org/TR/2010/REC-xpath20-20101214/> (besucht am 20.12.2011).



Burnard, Lou und Syd Bauman, Hrsg. (2014). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.7.0. Last updated on 16th September 2014, revision 13036. Charlottesville, Virginia: Text Encoding Initiative Consortium. URL: <http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf> (besucht am 17.09.2014).



Crockford, Douglas (2006). „JSON: The Fat-Free Alternative to XML“. In: *Proceedings of XML 2006 conference*. Boston. URL: <http://www.json.org/fatfree.html>
(besucht am 30.01.2012).

Bibliographie II

-  Dahlström, Erik, Patrick Dengler, Anthony Grasso, Chris Lilley, Cameron McCormack, Doug Schepers und Jonathan Watt (2011). *Scalable Vector Graphics (SVG) 1.1 (Second Edition)*. W3C Recommendation. World Wide Web Consortium (W3C). URL: <http://www.w3.org/TR/2011/REC-SVG11-20110816/>.
-  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.
-  ECMA International (2013). *The JSON Data Interchange Format*. ECMA Standard ECMA-404. Genf: ECMA International. URL: <http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf> (besucht am 10.10.2013).
-  Fricke, Ellen (2012). „Multimedialität und Multimodalität: Gesten und Lautsprache“. In: *Grammatik multimodal. Wie Wörter und Gesten zusammenwirken*. Bd. 40. Linguistik – Impulse und Tendenzen. Berlin und Boston: De Gruyter. Kap. 2, S. 36–75. DOI: 10.1515/9783110218893.36.

Bibliographie III

-  Grosso, Paul, Eve Maler, Jonathan Marsh und Norman Walsh (2003). *XPointer element() Scheme*. W3C Recommendation. World Wide Web Consortium (W3C). URL: <http://www.w3.org/TR/2003/REC-xptr-element-20030325/> (besucht am 20.12.2011).
-  Henschel, Renate (2003). *GeM Annotation Manual Version 2*. GeM project report 2003/04. Universität Bremen. URL: <http://www.fb10.uni-bremen.de/anglistik/langpro/projects/gem/corpus/Documentation/annotation-manual-2.pdf> (besucht am 24.09.2014).
-  Huitfeldt, Claus und C. M. Sperberg-McQueen (2004). „Markup Languages and Complex Documents (MLDC)“. Präsentation gehalten auf einer Kolloquiumsveranstaltung des Instituts der HKI (Historisch Kulturwissenschaftliche Informationsverarbeitung) der Universität zu Köln, 10. Dezember 2004. URL: <http://old.hki.uni-koeln.de/material/huitfeldt.ppt> (besucht am 02.04.2012).
-  Ide, Nancy M. (1998). „Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora“. In: *Proceedings of the First International Language Resources and Evaluation (LREC 1998)*. European Language Resources Association (ELRA). Granada, S. 463–470.
-  Ide, Nancy M., Greg Priest-Dorman und Jean Véronis (1996). *Corpus Encoding Standard (CES)*. Techn. Ber. Expert Advisory Group on Language Engineering Standards (EAGLES). URL: <http://www.cs.vassar.edu/CES/> (besucht am 20.12.2011).

Bibliographie IV

-  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.
-  Lücking, Andy und Thies Pfeiffer (2012). „Framing Multimodal Technical Communication. With Focal Points in Speech-Gesture-Integration and Gaze Recognition“. In: *Handbook of Technical Communication*. Hrsg. von Alexander Mehler und Laurent Romary. Bd. 8. Handbooks of Applied Linguistics. de Gruyter. Kap. 18, S. 591–644. DOI: 10.1515/9783110224948.
-  Malhotra, Ashok, Jim Melton, Norman Walsh und Michael Kay (2010). *XQuery 1.0 and XPath 2.0 Functions and Operators (Second Edition)*. W3C Recommendation. World Wide Web Consortium (W3C). URL: <http://www.w3.org/TR/2010/REC-xpath-functions-20101214/> (besucht am 20.12.2011).
-  Mann, William C. und Sandra A Thompson (1988). „Rhetorical structure theory: Toward a functional theory of text organization“. In: *Text – Interdisciplinary Journal for the Study of Discourse* 8.3, S. 243–281. DOI: 10.1515/text.1.1988.8.3.24.
-  Margareta, Eliza und Harald Lünen (2014). „Building Linguistic Corpora from Wikipedia Articles and Discussions“. In: *Journal for Language Technology and Computational Linguistics* 29.2.

Bibliographie V



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.



Marsh, Jonathan, David Orchard und Daniel Veillard (2006). *XML Inclusions (XInclude) Version 1.0 (Second Edition)*. W3C Recommendation. World Wide Web Consortium (W3C). URL: <http://www.w3.org/TR/2006/REC-xinclude-20061115/> (besucht am 20.12.2011).



Martinec, Radan und Andrew Salway (2005). „A system for image–text relations in new (and old) media“. In: *Visual Communication* 4.3, S. 339–374.



Nelson, Theodor Holm (1997). „Embedded Markup Considered Harmful“. In: *XML. Principles, Tools and Techniques*. Hrsg. von Dan Connolly. Bd. 2. The World Wide Web Journal 4. O'Reilly Media, S. 129–134. URL: <http://www.xml.com/pub/a/w3j/s3.nelson.html> (besucht am 20.12.2011).



Proceedings of Balisage: The Markup Conference (2008). Bd. 1. Balisage Series on Markup Technologies. Montréal.



Proceedings of Balisage: The Markup Conference (2009). Bd. 3. Balisage Series on Markup Technologies. Montréal.



Sperberg-McQueen, C. M. und Lou Burnard, Hrsg. (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.

Bibliographie VI

-  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. URL: <http://llc.oxfordjournals.org/content/14/1/29.full.pdf> (besucht am 07.04.2012).
-  — (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.
-  — (2008). „Markup Discontinued Discontinuity in TexMecs, Goddag structures, and rabbit/duck grammars“. In: *Proceedings of Balisage: The Markup Conference*. Bd. 1. Balisage Series on Markup Technologies. Montréal. DOI: 10.4242/BalisageVol1.Sperberg-McQueen01.
-  Stegmann, Jens und Andreas Witt (2009). „TEI Feature Structures as a Representation Format for Multiple Annotation and Generic XML Documents“. In: *Proceedings of Balisage: The Markup Conference*. Bd. 3. Balisage Series on Markup Technologies. Montréal. DOI: 10.4242/BalisageVol3.Stegmann01.

Bibliographie VII

-  Stöckl, Hartmut (2008). „Sprache-Bild-Bezüge im Text“. In: *Die Sprache im Bild – Das Bild in der Sprache. Zur Verknüpfung von Sprache und Bild im massenmedialen Text: Konzepte. Theorien. Analysemethoden.* Linguistik – Impulse und Tendenzen 3. Berlin und Boston: De Gruyter. Kap. 5, S. 242–300.
-  — (2009). „Beyond Depicting. Language-Image-Links in the Service of Advertising“. In: *AAA. Arbeiten aus Anglistik und Amerikanistik* 34.1, S. 3–28.
-  — (2011). „Sprache-Bild-Texte lesen. Bausteine zur Methodik einer Grundkompetenz“. In: *Bildlinguistik. Theorien – Methoden – Fallbeispiele.* Hrsg. von Hajo Diemannshenke, Michael Klemm und Hartmut Stöckl. Berlin: Erich Schmidt Verlag, S. 43–70.
-  Stührenberg, Maik (2012). „Auszeichnungssprachen für linguistische Korpora: Theroretische Grundlagen, De-facto-Standards, Normen“. Dissertation. Universität Bielefeld. URL: <http://pub.uni-bielefeld.de/publication/2492772> (besucht am 30.04.2012).
-  — (2013a). „Quo vadis XML? History and possible future directions of the Extensible Markup Language“. In: *XML Prague 2013 Conference Proceedings.* Hrsg. von Jirka Kosek. University of Economics. Prag, S. 141–162. URL: <http://archive.xmlprague.cz/2013/files/xmlprague-2013-proceedings.pdf> (besucht am 08.02.2013).

Bibliographie VIII



Stührenberg, Maik (2013b). „What, when, where? Spatial and temporal annotations with XStandoff“. In: *Proceedings of Balisage: The Markup Conference*. Bd. 10. Balisage Series on Markup Technologies. Montréal. DOI: 10.4242/BalisageVol10.Stuhrenberg01.



— (2014). „Less Destructive Cleaning of Web Documents by Using Standoff Annotation“. In: *Proceedings of the 9th Web as Corpus Workshop (WaC-9). Held in conjunction with the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Hrsg. von Felix Bildhauer und Roland Schäfer. Gothenburg, Sweden: Association for Computational Linguistics, S. 16–21. URL: <http://www.aclweb.org/anthology/W14-0403> (besucht am 26.04.2014).



Stührenberg, Maik und Daniela Goecke (2008). „SGF – An integrated model for multiple annotations and its application in a linguistic domain“. In: *Proceedings of Balisage: The Markup Conference*. Bd. 1. Balisage Series on Markup Technologies. Montréal. DOI: 10.4242/BalisageVol1.Stuehrenberg01.

Bibliographie IX

-  Stührenberg, Maik, Daniela Goecke, Nils Diewald, Irene Cramer und Alexander Mehler (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. DOI: 10.3115/1642059.1642082. URL: <http://www.aclweb.org/anthology/W/W07/W07-1523.pdf> (besucht am 22.05.2012).
-  Stührenberg, Maik und Daniel Jettka (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/BalisageVol13.Stuhrenberg01.
-  Van Leeuwen, Theo (2005). *Introducing Social Semiotics: An Introductory Textbook*. Routledge Chapman & Hall. Kap. Information Linking.
-  von Engelhardt, Jörg (2002). „The Language of Graphics“. Diss. University of Amsterdam. URL: <http://dare.uva.nl/record/105970>.
-  Witt, Andreas, Daniela Goecke, Felix Sasaki und Harald Lungen (2005). „Unification of XML Documents with Concurrent Markup“. In: *Literary and Linguistic Computing* 20.1, S. 103–116.