*Manuela Kunze, Dietmar Rösner*

# Die Document Suite XDOC: ein Fortschrittsbericht

Zusammenfassung: Die aktuellen Entwicklungen zur Aufbereitung von Dokumenten – z.B. für das Semantic Web – und der technologische Fortschritt erfordern einen schnellen und problembezogenen Zugriff auf Informationen in Dokumentbeständen.

In diesem Beitrag wird der aktuelle Stand der Document Suite XDOC vorgestellt. XDOC nutzt Methoden der Computerlinguistik und der Wissensverarbeitung, um Voraussetzungen für den Umgang mit großen Dokumentbeständen, z.B. für effektives Information Retrieval (IR) und Informationsextraktion (IE) zu schaffen.

## 1 Einleitung

In XDOC werden verschiedene Methoden für die linguistische Verarbeitung deutschsprachiger Dokumente vereint, mit dem Ziel, die in natürlichsprachlicher Form vorliegenden Informationen in den Dokumenten für weitergehende Auswertungen (z.B. Applikationen im Bereich des IR und IE) aufzubereiten.

Die Document Suite wird u.a. als ein Werkzeug für die rechtsmedizinische Forschung anhand von Obduktionsprotokollen eingesetzt. Eine Evaluierung der Methoden von XDOC erfolgt derzeit im Rahmen einer Zusammenarbeit mit dem Institut für Rechtsmedizin der Universität Magdeburg im Bereich der Verkehrsunfallforschung. Ein konkretes Anwendungsszenario kann wie folgt beschrieben werden: In einem großen Bestand von Obduktionsprotokollen sollen automatisch anhand von medizinischen Merkmalen (z.B. Verletzungen in der Kniegegend) sowie anhand situationsbedingter Parameter (z.B. Airbag, Gurt) Protokolle für detaillierte Analysen ausgewählt werden. Anhand dieses Materials sollen dann typische Verletzungsmuster bei unterschiedlichen Unfallhergängen erkannt werden.

Weitere Einsatzgebiete der Document Suite XDOC sind:

- die Informationsextraktion aus technischer Dokumentation und
- die automatische Generierung von strukturierten Firmenprofilen aus den Internetpräsentationen von Firmen.

In den nachfolgenden Abschnitten werden kurz die verschiedenen in XDOC integrierten Methoden beschrieben und auf einige charakteristische Merkmale der Methoden eingegangen.

## 2 Die Document Suite XDOC

In XDOC werden für die Extraktion und Auszeichnung von inhaltsbezogenen Informationen linguistische Methoden kombiniert. Eine funktionelle Einteilung der zur Verfügung stehenden Methoden ist in Abbildung 1 zu sehen. In den nachfolgenden Abschnitten werden die Methoden der verschiedenen Module vorgestellt.

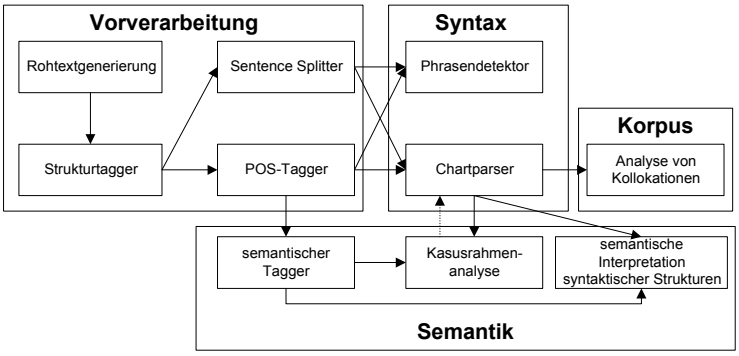


Abb. 1: Architektur von XDOC.

### 2.1 Vorverarbeitung

In diesem Modul werden die Methoden vereint, die essenziell für die nachfolgenden Module sind. Schwerpunkt des Moduls ist die Transformation der zu analysierenden Dokumente aus den verschiedenen Ausgangsformaten in ein kanonisches verarbeitbares Textformat (s. Abb. 2). Für diesen Zweck beinhaltet das Modul zur Vorverarbeitung einen **Rohtextgenerierer**. Die Aufgabe des Rohtextgenerierers ist die Abbildung von Dokumenten eines speziellen Formats (z.B. RTF) in ein kanonisches XML-basiertes Layout. Dabei werden die layoutbezogenen Informationen (z. B. Schriftart, Absatzformat) als Attribute in das XML-Dokument übernommen. Bei bekannten Dokumentstrukturen kann auch eine logische Segmentierung des Textes erfolgen aufgrund von speziellen Token (z.B. „Innere Besichtigung“), die logische Einheiten einleiten, oder auch aufgrund von Positions- und Layoutmerkmale.

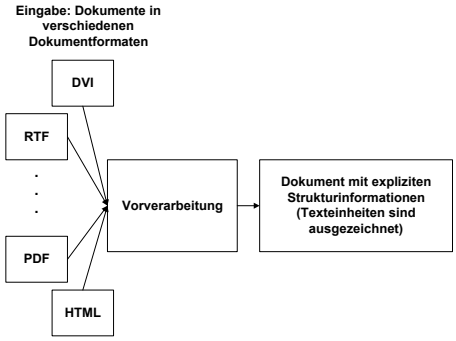


Abb. 2: Integration verschiedener Dokumentvorlagen.

Weiterhin ist in diesem Modul ein **Strukturtagger** integriert. Der Strukturerkenner realisiert die Annotation von u.a. Abkürzungen, Initialen und Interpunktion im Text. Für die Erkennung von Abkürzungen wird ein Lexikon benutzt, in dem allgemeine Abkürzungen der jeweiligen Sprache des zu verarbeitenden Dokuments codiert sind. [1] Die annotierten Interpunktionen werden vom **Satzerkenner** genutzt, um potentielle Satzstrukturen auszuzeichnen, da nachfolgende Methoden (wie z.B. der syntaktische Parser) als Eingabe Satzstrukturen erwarten.

<SATZ><PRP>Am</PRP> <ADJ>linken</ADJ> <N>Handgelenk</N> <N SRC="UNG"> Abblassung</N> <IP>(</IP> <XXX>vermutlich</XXX> <N SRC="UC1">Armbanduhr</N> <IP>)</IP> <IP>.</IP></SATZ>
<SATZ><ORD>18.</ORD> <N>Beine</N> <MULT VAL="PART ADV">nicht</MULT> <ADV>sehr</ADV> <XXX>muskelkräftig</XXX><IP>.</IP></SATZ>

Bsp. 1: Text mit Ergebnissen des Satztrenners und POS-Tagger.

Als Vorstufe zur weiteren syntaktischen und semantischen Verarbeitung von textuellen Dokumenten ist es sinnvoll, den vorgefundenen, zusammenhängenden Zeichenketten – den Token – zunächst Wortklassen zuzuordnen (sog. part-of-speech- oder **POS-Tagging**). Dabei kann es sich sowohl um „klassische" Wortklassen aus der Linguistik handeln (z.B. Nomen, Verb, Adjektiv, Artikel, Präposition und andere sog. lexikalische Kategorien) wie auch um nicht-lexikalische Kategorien, die dann meist eine gebietsspezifische Relevanz haben (z.B. Telefonnummern, E-Mail-Adressen, Materialkennungen, Substanzbezeichner, ...). Für die Verarbeitung deutscher Texte nutzt der POS-Tagger die Morphologiekomponente MORPHIX (Finkler et. al. 1988). Diese

[1] Experimentiell wurden die Methoden von XDOC auch für die Verarbeitung von englischsprachigen Dokumenten genutzt. Dies wurde durch ein Austauschen der Ressourcen verschiedener Methoden realisiert (Kunze et.al. 2002).

Komponente zeichnet sich dadurch aus, dass die *geschlossenen* Wortklassen des Deutschen (d.h. Artikel, Präpositionen, Pronomina, Konjunktionen, usw.) sowie alle unregelmäßigen Verben in allen ihren Wortformen vollständig abgedeckt sind. Für die *offenen* Wortklassen (also Nomen, Verben, Adjektive, Adverbien) sind die Muster der möglichen Formen (linguistisch: Flexionsparadigmen) abgedeckt.

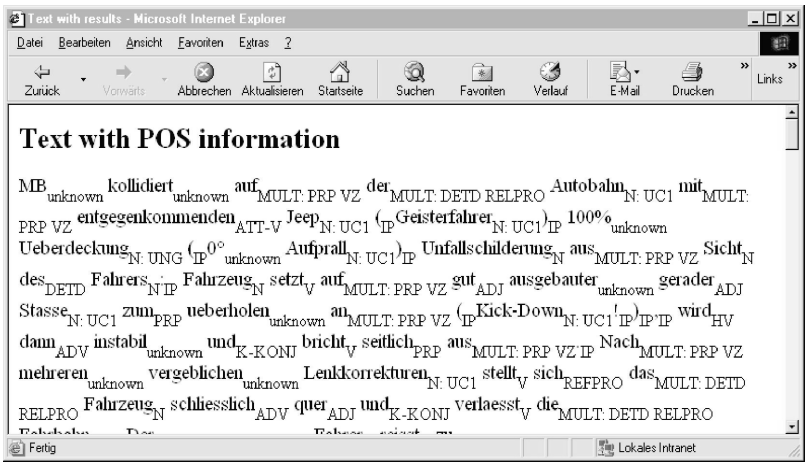


Abb. 3: Text mit POS-Informationen.[2]

## 2.2 Syntaktisches Modul

Im syntaktischen Modul ist ein flexibler Bottom-up Chartparser eingebunden. Dieser kann kontextfreie Grammatikregeln. Auf der rechten Seite sind auch Kombinationen aus Kategorien und terminalen Symbolen (d.h. Wörtern) möglich. Damit sind phrasale Muster und nicht nur Phrasenstrukturregeln (in der Art S→NP VP) darstell- und verarbeitbar.

Der **syntaktische Chartparser** erwartet als Eingabe einen mit POS-Informationen ausgezeichneten Text. Das Ergebnis der syntaktischen Analyse liefert unter Umständen mehrere Lesarten für einen Satz oder eine Phrase. Diese Ambiguitäten können teilweise über die nachfolgende semantische Analyse aufgelöst werden.

[2] Hervorhebungen werden in der Web-Schnittstelle von XDOC farblich realisiert. Für eine bessere Lesbarkeit wurden in diesem Beispiel stattdessen als Markierung Fettdruck verwendet bzw. die relevanten Informationen wurden „tiefgestellt“.

Für das syntaktische Parsing wird eine Grammatik benutzt, die kontextfreie Regeln enthalten, die mit Merkmalsstrukturen annotiert sind. Der Vorteil eines Chart-Parsers ist, dass die Mehrfachanalyse bereits erfolgreich verarbeiteter Teilstrukturen vermieden wird und alle – auch partiellen – Analyseergebnisse in einer kompakten Datenstruktur erhalten bleiben (siehe z.B. (Gazdar et. al. 1989)).

Die in der Document Suite verwendeten Grammatiken sind modular organisiert. Gruppen von Regeln können flexibel aktiviert und deaktiviert werden. Dies wird ausgenutzt, um den Konventionen unterschiedlicher Subsprachen Rechnung tragen zu können. Je nach Gebiet können Subsprachen syntaktische Konstruktionen präferieren, die in der Standardsprache ungewöhnlich oder gar ungrammatisch sind.

Eine ausführliche Beschreibung der Methoden des syntaktischen Moduls von XDOC ist in Rösner et. al. 2002a zu finden.

Phrasale Muster findet man in allen Dokumentbeständen. In den Webseiten von produzierenden Firmen sind häufig Formulierungen wie „unsere Kunden sind ...“, „unsere Abnehmer sind ...“ oder „wir produzieren für ...“ zu finden. Diese Muster kann man verallgemeinern und in Form von phrasalen Regeln beschreiben, die dann vom Parser ausgewertet werden. Mit dem in XDOC integrierten Chartparser ist es möglich, die Grammatik durch eine Liste von phrasalen Mustern zu ersetzen und den Chartparser so als **Phrasendetektor** zu nutzen. Auf diesem Wege kann eine schnelle erste Klassifizierung von Dokumenten vorgenommen werden. In Abbildung 4 ist ein Ausschnitt aus einem Obduktionsprotokoll zu sehen. In diesem Textausschnitt wurden die erkannten pathologischen Merkmale hervorgehoben, die für eine Klassifizierung relevant sind.

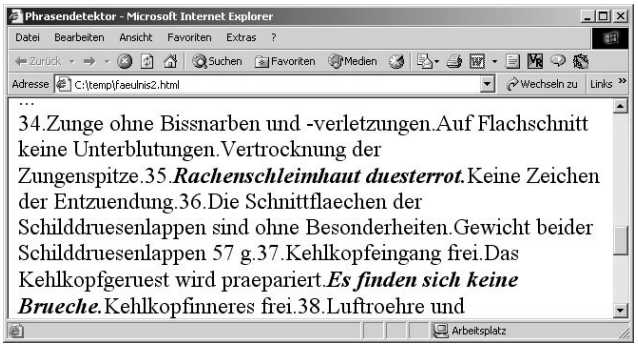


Abb. 4: Hervorhebung von für die Klassifizierung relevanten Merkmalen.

## 2.3 Semantisches Modul

Das semantische Modul von XDOC beinhaltet einen **semantischen Tagger.** Dieser weist jedem Token die zugehörige semantische Kategorie zu. Dafür wird ein Lexikon benutzt, welches neben der semantischen Kategorie des Tokens auch den zugehörigen Kasusrahmen (bei Nomen und Verben) enthält. Folgende Ergebnisse können auftreten: Einem Token kann eine eindeutige semantische Interpretation zugeordnet werden, es sind mehrere semantische Interpretationen des Tokens möglich bzw. das Token kann keiner semantischen Rolle zugeordnet werden, da der zugehörige Lexikoneintrag fehlt.

In Abbildung 5 wurden in einem Obduktionsprotokoll diejenigen Token hervorgehoben, die aufgrund der Analyse des semantischen Taggers mögliche Verletzungen beschreiben.

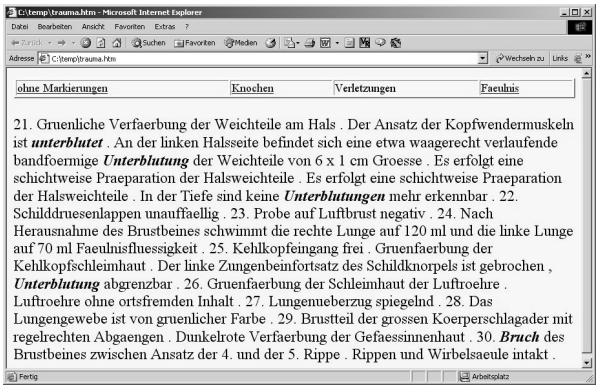


Abb. 5: Hervorhebung von semantischen Kategorien (Verletzungen).

Neben dem semantischen Tagger beinhaltet das semantische Modul eine **Kasusrahmenanalyse**. Die Kasusrahmenanalyse erwartet als Eingabe einen mit semantischen Tags ausgezeichneten Text und die Ergebnisse des syntaktischen Parsers. Durch die Auswertung der Informationen des Kasusrahmens werden komplexe Konzepte innerhalb einer linguistischen Struktur erkannt. Dabei werden Relationen zwischen den vom semantischen Tagger erkannten isoliert stehenden Konzepten anhand der Informationen des Kasusrahmens zugeordnet. Der Kasusrahmen enthält neben der semantischen Kategorie des Wortes auch die Beschreibung der syntaktischen und semantischen Anforderungen an die Füller der möglichen Relationen.

Weiterhin beinhaltet das semantische Modul die **semantische Interpretation von spezifischen syntaktischen Strukturen**. Einige Strukturen der deutschen Syntax können unabhängig von den Token in der Struktur semantisch gleich interpretiert werden, z.B. können Adjektive Nomen genauer spezifizieren, während Adverbien

Prozesse (Verben) näher beschreiben. Ein anderes Beispiel ist der Gleichstellungsnominativ, der als taxonomische Aussage interpretiert werden kann, oder bestimmte Formulierungen beginnend mit *als* („*Als formlose Stoffe werden Gase, Pulver, Flüssigkeiten bezeichnet.*“).

In Subsprachen, insbesondere in Befundformulierungen, werden kurze syntaktische Strukturen häufig genutzt: Erste Analysen mit einem Korpus von 400 Obduktionsprotokollen zeigten, dass ca. 40 Prozent der „Sätze“ aus dem Befundteil einfache Strukturen der Form NP ADJP|V (z.B.: „Niere unauffällig.“) oder der Form PP NP (z.B. „In der Gallenblase reichlich Flüssigkeit.“) sind. Diese Strukturen und ihre verallgemeinerte semantische Interpretation wurden in einem Lexikon codiert, so dass es möglich ist, die Ergebnisse des syntaktischen Parsers auf einfachem Wege in eine semantische Interpretation zu überführen. Experimentell wurde dafür zunächst der Topic Map Ansatz (siehe Pepper et. al. 2001) gewählt, je nach Anwendung kann hier aber auch eine andere Präsentationsform genommen werden.

### 2.4 Korpusbasiertes Modul

Dieses Modul beinhaltet im momentanen Entwicklungsstand die Analyse von Kollokationen. Als Kollokationen werden in der Document Suite XDOC nicht nur die Kombinationen von einzelnen Token betrachtet, sondern Kombinationen syntaktischer Strukturen. Es wird z.B. dabei untersucht, welche Kombinationen von NPs mit Adjektiven oder Verben bzw. auch PPs innerhalb des Korpus möglich sind. Dabei werden auch Negatoren (z.B. „Arme nicht muskelkräftig.“) und Modifikatoren (z.B. „Augen spaltweit geöffnet.“) berücksichtigt.

Die Ergebnisse dieser Analyse werden für die Erstellung initialer Ontologien und Lexika genutzt. So können Synonyme und Paraphrasen erkannt sowie auch Konzeptnamen der jeweiligen Domäne gelernt werden (siehe (Rösner et. al. 2002b)).

## 3 Designrichtlinien

Für alle Anwendungen/Tools im Bereich der natürlichen Sprachverarbeitung sind u.a. folgende Schwerpunkte bei der Entwicklung und Design der Methoden zu berücksichtigen:

- Annotierung: Bei der Sprachverarbeitung ist es erforderlich unterschiedliche Strukturen auszuzeichnen (von linearen Strukturen bis hin zu komplexen Baum-

strukturen). Wünschenswert ist also ein Formalismus, mit dem es möglich ist, verschiedenartige Strukturen auf einheitliche Weise auszuzeichnen.
- Robustheit: Durch die Produktivität der natürlichen Sprache müssen die Methoden in der Lage sein, lexikalische Lücken abzudecken bzw. damit umgehen zu können.
- Ambiguitäten: Bei der Sprachverarbeitung treten schon auf kleinster Ebene (z.B. Tokenizer) Ambiguitäten auf, die (mitunter) auf den höheren Ebenen aufgelöst werden können. Es ist aber auch möglich, dass diese Ambiguitäten auf den höheren Ebenen der Sprachverarbeitung weiterhin bestehen können. Die Methoden sollten dafür eine geeignete Strategie unterstützen, um mit solchen Ambiguitäten umgehen zu können.

Nach der kurzen Vorstellung der einzelnen Methoden von XDOC im Kapitel 2, wird jetzt in Hinblick auf die zuvor beschriebenen Kriterien auf die Merkmale einiger implementierter Methoden eingegangen.

### 3.1 Einheitlicher Beschreibungsformalismus von Daten

Die Extensible Markup Language (XML) ist eine Metasprache, mit der es möglich ist ein eigenes – und somit ein anwendungsbezogenes – Markup zu definieren. XML (siehe Bray et. al. 2000) ist vereinfachtes SGML, aber strukturierter und leistungsfähiger als HTML und in Bezug auf die Document Suite ein optimaler Formalismus zur Beschreibung der verschiedenen Ergebnisstrukturen. Des Weiteren wurden vom W3C weitere Spezifikationen bzw. Bestrebungen erarbeitet um Formalismen in Zusammenhang mit XML zu definieren, z.B. XSL. XSL kann als eine eigenständige funktionale Sprache betrachtet werden, es ist mehr als nur ein Stylesheet. Durch XSL Transformation (XSLT) können die Ergebnisse von XDOC benutzer- und problembezogen aufbereitet werden. Für den Endnutzer ist es z.B. sinnvoll, wenn die Ergebnisse (z.B. im Dokument relevante Stellen) farblich hervorgehoben werden bzw. der Browser (als Visualisierungswerkzeug von Dokumenten) zu den betreffenden Stellen im Dokument navigiert, währenddessen sich der NLP-Experte sich für detaillierte (Teil-) Ergebnisse (z.B. bei der Grammatikentwicklung: die Erkennung von Features syntaktischer Strukturen) interessiert. Des Weiteren ist es möglich, mit XSL weitergehende Auswertungen durchzuführen, z.B. statistische Auswertungen über einen Dokumentenkorpus. Das Ausgabeformat der Transformationen ist ebenso flexibel – innerhalb von XDOC werden das HTML- und das RTF-Format unterstützt (siehe Abbildung 6).

Ein weiterer Vorteil der Nutzung von XML ist die Möglichkeit, verwendete XML-Strukturen – Ressourcen bzw. auch Ergebnisse – in Form von DTDs oder auch durch XML Schemata zu beschreiben. Ein Import der Ergebnisse in andere Anwendungen gestaltet sich somit einfacher.

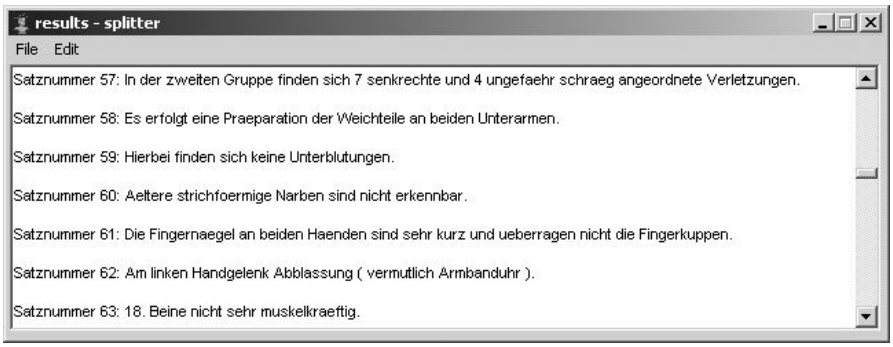


Abb. 6: Ergebnisse des Satztrenners (aus Beispiel 1) als RTF-Dokument aufbereitet.

Innerhalb von XDOC wird XML für folgende Zwecke genutzt:

- Eingaben werden als XML-Strukturen erwartet,
- Ergebnisse werden ebenfalls als XML-Strukturen ausgegeben und
- alle Ressourcen sind in XML kodiert.

Für die Ergebnisstrukturen und Ressourcenbeschreibung wurden DTDs spezifiziert, um einen einfachen Zugang zu den Daten zu realisieren. Nachfolgend ist die DTD zur Beschreibung der Ergebnisstruktur der Kasusrahmenanalyse angegeben:

```
<!ELEMENT CONCEPTS CONCEPT*>
    <!ELEMENT CONCEPT (WORD, DESC, SLOTS?)>
        <!ATTLIST CONCEPT TYPE CDATA #REQUIRED>
      <!ELEMENT WORD (#PCDATA)>
      <!ELEMENT DESC (#PCDATA)>
      <!ELEMENT SLOTS (RELATION+)>
        <!ELEMENT RELATION (ASSIGN_TO, FORM, CONTENT)>
                    <!ATTLIST RELATION TYPE CDATA #REQUIRED>
            <!ELEMENT ASSIGN_TO (#PCDATA)>
            <!ELEMENT FORM (#PCDATA)>
            <!ELEMENT CONTENT (#PCDATA)>
```

Bsp. 2: DTD für die Ressource der Kasusrahmenanalyse.

Wie schon im Kapitel 2.1 beschrieben, werden Dokumentbeschreibungen (XML-Schema) auch für die Gewinnung von Textabschnitten aus dem ursprünglichen Dokumentformat (z.B. Word) benutzt. In Abbildung 7 ist ein automatisch strukturell ausgezeichnetes Dokument – in diesem Fall handelt es sich um ein Obduktionsprotokoll – zu sehen.


Abb. 7: Strukturell ausgezeichnetes Obduktionsprotokoll.

## 3.2 Robustheit

Viele Methoden von XDOC arbeiten mit sprach- und domänenbezogenen Lexika. Aufgrund der Produktivität der Sprache und von domänenspezifischen Fachtermini kann nicht davon ausgegangen werden, dass mit den Ressourcen eine vollständige Abdeckung realisiert werden kann. Deshalb wurde innerhalb von XDOC auf Robustheit im Umgang mit lexikalischen und konzeptuellen Lücken besonderer Wert gelegt.

### 3.2.1 Beispiel: POS-Tagger

Da nicht erwartet werden kann, dass das Lexikon jemals vollständig ist (z.B. Neologismen, produktive Kompositabildung, Fachtermini), muss mit der Situation umgegangen werden können, dass für einige Token keine MORPHIX-Analyse gefunden wird.

Hier werden zwei Techniken eingesetzt: Zunächst wird versucht, Heuristiken anzuwenden, die sich auf solche Aspekte des Token stützen, die mit Stringanalyse einfach festgestellt werden können (z.B. Groß-/Kleinschreibung, Endungen, ...), und welche zusätzlich ggf. die relative Position des Token in Bezug auf eine Satzgrenze mit einbeziehen. Führt dies zu einer Klassifikation, so wird das Token mit dem Bezeichner der zugehörigen POS-Klasse markiert, die verwendete Heuristik wird als

Wert des Attribut SRC angegeben (vgl. Beispiel 3). Wenn keine Heuristik anwendbar ist, wird das Token als Element der Klasse 'unbekannt' eingestuft (Tag „XXX").

```
<NP TYPE="COMPLEX" RULE="NPC3" GEN="FEM" NUM="PL" CAS="_">
  <NP TYPE="FULL" RULE="NP1" CAS="_" NUM="PL" GEN="FEM">
    <N SRC="UNG">Ablagerungen</N>
  </NP>
  <PP CAS="DAT">
    <PRP CAS="DAT">an</PRP>
    <NP TYPE="FULL" RULE="NP2" CAS="DAT" NUM="SG" GEN="FEM">
      <DETD>der</DETD>
      <N SRC="UC1">Blutader</N>
    </NP>
  </PP>
</NP>
```

Bsp. 3: Unbekannte Token mit Heuristik als Nomen klassifiziert.

### 3.2.2 Beispiel: Semantischer Tagger

Folgende Ergebnisse können beim semantischen Tagger erwartet werden:

- Es gibt eine eindeutige Zuordnung, d.h. nur eine Lesart/Interpretation ist möglich.
- Es sind mehrere semantische Lesarten möglich. Ähnlich zum POS-Tagger werden alle Ergebnisse ausgegeben, d.h. das Attribut des Elements beinhaltet mehrere semantische Rollen des Tokens. In diesem Fall erfolgt die Disambiguierung durch die nachgelagerte Kasusrahmenanalyse.
- Das Wort ist morphologisch und semantisch unbekannt. In diesem Fall erfolgt lediglich eine Auszeichnung mit dem Tag „XXX".
- Das Wort konnte einer Wortklasse zugeordnet werden, ist aber semantisch gesehen ein unbekanntes Token (d.h. der dazugehörige Eintrag im semantischen Lexikon fehlt). In diesem Fall erfolgt eine kanonische Zuordnung anhand der erkannten POS-Informationen. Die Zuordnung sieht folgendermaßen aus:
  - Nomen werden als semantische Konzepte („CONCEPT" ohne genaue Angabe des Typs, d.h. ohne Angabe eines Attributes) aufgefasst.
  - Konjunktionen werden als semantische/logische Verknüpfungen interpretiert.
  - Adjektive werden als Eigenschaften („PROPERTY") interpretiert.
  - Sowohl Verben als auch Präpositionen beschreiben semantisch betrachtet, Relationen zwischen Konzepten. Verben und Präpositionen werden somit als 'RELATION' bzw. 'PRELATION' beschrieben.

### 3.3 Auflösung von Ambiguitäten

Ein Problem bei computerlinguistischen Auswertungen ist die Auflösung von Ambiguitäten. Ambiguitäten werden in XDOC i.d.R. in den nachfolgenden Prozessschritten aufgelöst. In Tabelle 1 erfolgt eine Auflistung der Prozessschritte, bei denen Ambiguitäten zu erwarten sind und es wird angegeben, durch welche nachgelagerten Prozesse eine Auflösung dieser Ambiguitäten realisiert wird.

| **Auftretende Ambiguitäten in ...** | **Auflösung durch ...** |
|---|---|
| POS Tagger | Syntaktischen Parser |
| Syntaktischer Parser | Kasusrahmenanalyse |
| Semantischer Tagger | Kasusrahmenanalyse |
| Phrasendetektor | Auswertung der Ergebnisse des Phrasendetektors |

Tabelle 1: Auflösung von Ambiguitäten.

In den nachfolgenden Kapiteln wird die Auflösung von auftretenden Ambiguitäten auf morphologischer (POS Tagger) und semantischer Ebene (semantischer Tagger) beschrieben.

#### 3.3.1 Beispiel: Syntaktischer Parser

Im Abschnitt 3.2.1 wurden die möglichen Ergebnisse des POS-Taggers beschrieben. Neben der eindeutigen Klassifizierung eines Tokens (zu einer Wortklasse bzw. die Einstufung als „unbekannt“) ist es möglich, dass für ein Token verschiedene Klassifizierungen erfolgen können. Im Beispiel 4 wurden die Token des Satzes „Die weichen Hirnhaeute sind zart und durchscheinend.“ mit POS-Informationen versehen.

```
<MULT VAL="DETD RELPRO">Die</MULT> <MULT VAL="V ADJ">weichen</MULT> <N>Hirnhaeute</N> <HV>sind</HV> <ADJ>weich</ADJ> <K-KONJ>und</K-KONJ> <XXX> durchscheinend</XXX><IP>.</IP>
```

Bsp. 4: Ergebnis des POS-Taggers.

Eine ambige Zuordnung erfolgte für die Token „die“ (Determiner, Relativpronomen) und „weich“ (Verb, Adjektiv) – erkennbar durch den Tag „MULT“. Für das Token „durchscheinend“ war keine Einordnung möglich, es wurde als ein unbekanntes Token ausgezeichnet (Tag „XXX“).

Bei der weiteren Auswertung des Satzes durch den syntaktischen Parser kann eine Auflösung der Ambiguitäten der beiden Token erfolgen. Der Parser findet eine syntaktische Lesart für den Satz (siehe Beispiel 5), bei der das Token „die“ als Determiner

und das Token „weich“ als Adjektiv (durch die Regel „NP2“) interpretiert werden, da eine Regel gefunden wurde, die eine Interpretation der Token zulässt. Weiterhin wird bei dem Parser-Ergebnis für das unbekannte Token „durchscheinend“ angenommen, dass es sich hierbei um ein Adjektiv handelt (durch die Regel „MA13“).

```
<S TYPE="MEDICAL-ATTRIBUTE-VALUE-SENTENCE" RULE="NPMAVS1" CAS="_">
 <MA TYPE="MEDICAL-ATTRIBUTE" RULE="MA27" CAS="_">
  <NP TYPE="FULL" RULE="NP2" CAS="NOM" NUM="PL" GEN="FEM">
   <DETD>Die</DETD>
   <ADJ>weichen</ADJ>
   <N>Hirnhaeute</N>
  </NP>
  <HV ROOT="sei" FLEX="FIN">sind</HV>
  <ADJ>weich</ADJ>
  <K-KONJ>und</K-KONJ>
  <XXX AS="ADJ">durchscheinend</XXX>
 </MA>
 <IP>.</IP>
</S>
```

Bsp. 5: Disambiguierung durch den syntaktischen Parser.

### 3.3.2 Beispiel: Kasusrahmenanalyse

Ein Problem beim semantischen Tagger ist die Disambiguierung unterschiedlicher Wortbedeutungen. In XDOC wird dafür die Kasusrahmenanalyse genutzt. Im Beispiel 6 ist ein Auszug von Ergebnissen des semantischen Taggers zu sehen, es handelt sich hierbei um die Auszeichnung der Phrase „Fertigen fester Koerper aus formlosem Stoff“. Das Token „Stoff“ kann hierbei nicht eindeutig zugeordnet werden.

```
<CONCEPT TYPE="process">Fertigen</CONCEPT> <PROPERTY TYPE="state">fester</PROPERTY>
<CONCEPT TYPE="object">Koerper</CONCEPT> <PRELATION TYPE="herkunft">aus
</PRELATION> <PROPERTY TYPE="state ">formlosem</PROPERTY> <CONCEPT TYPE="object
material">Stoff</CONCEPT>
```

Bsp. 6: Ergebnis des semantischen Taggers.

Durch die nachgelagerte Analyse des Kasusrahmens des Tokens „Fertigen“ kann hier eine Auflösung erfolgen (siehe Beispiel 7). Das Token „Stoff“ kann nur dann einem Rollenfüller zugeordnet werden, wenn es der semantischen Interpretation vom Typ „Material“ entspricht (Tag „SOURCE“).

```
<CONCEPT TYPE="process">
        <WORD>Fertigen</WORD>
        <DESC>Schaffung von etwas</DESC>
        <SLOTS>
                <RELATION TYPE="RESULT">
                        <ASSIGN_TO>OBJECT</ASSIGN_TO>
                        <FORM>N(gen, fak) P(akk, fak, von)</FORM>
                        <CONTENT>fester Koerper</CONTENT>
                </RELATION>
                <RELATION TYPE="SOURCE">
                        <ASSIGN_TO>MATERIAL</ASSIGN_TO>
                        <FORM>P(dat, fak, aus)</FORM>
                        <CONTENT>aus formlosem Stoff</CONTENT>
                </RELATION>
        </SLOTS>
</CONCEPT>
```

Bsp. 7: Disambiguierung durch die Kasusrahmenanalyse.

## 4 Zusammenfassung

In diesem Beitrag wurde der aktuelle Stand der Document Suite XDOC – eine Sammlung von Methoden für eine flexible und robuste Verarbeitung von Dokumenten auf der Basis von XML – vorgestellt. Alle Methoden in XDOC sind domänenunabhängig implementiert, bzgl. der verwendeten Ressourcen kann es jedoch erforderlich sein, domänenspezifische Terminologie bzw. phrasale Muster zu ergänzen. Ähnlich dem GATE-System (Cunningham et. al. 2002) ist XDOC ebenfalls modular aufgebaut. Dadurch ist es möglich, komplexe Funktionalitäten für die Analyse von Dokumenten zur Verfügung zu stellen.

Der Schwerpunkt zukünftiger Arbeiten liegt in der Auflösung von Anaphern (z.B. „Dunsung derselben.“) und bei der Erkennung und Auflösung von koordinierten Strukturen (z.B. „Augenunter- und –oberlid“). Da sich das zur Verfügung stehende Material ständig vergrößert, ist die Erweiterung der korpusbasierten Methoden um Lernverfahren ein wichtiger Schritt (z.B. für die korpusbasierte Grammatikentwicklung).

## Literatur

Bray, T./ Paoli, J./ Sperberg-McQueen, C.M./ Maler, E.(eds.)(2000): Extensible Markup Language (XML) 1.0 (second Edition). World Wide Web Konsortium, W3C Recommendation. URL:http//www.w3.org/TR/2000/REC-xml-20001006. 2000.

Cunningham, Hamish/ Wilks, Yorick (1996): GATE – a General Architecture for Text Engineering. In: Computing and Humanities, Vol. 36, 2002, 223-254.

Finkler, Wolfgang/ Neumann, Günter (1988): MORPHIX: A fast Realization of a classification-based Approach to Morphology. In: Proceedings der 4. Österreichischen Artificial-Intelligence Tagung, Wiener Workshop Wissensbasierte Sprachverarbeitung, 11-19.

Gazdar, Gerald/ Mellish, Chris (1989): Natural Language Processing in LISP: An Introduction to Computational Linguistics. Addison-Wesley.

Kunze, Manuela/ Xiao, Chun (2002): An Approach for Resource Sharing in Multilingual NLP. In: Proceedings of STarting Artificial Intelligence Researchers Symposium STAIRS. 2002, 123-124.

Pepper, Steve/ Moore, Graham(eds.) (2001): XML Topic Maps (XTM) 1.0. Specification, TopicMaps.Org. http://www.topicmaps.org/xtm/1.0/. 2001.

Rösner, Dietmar/ Kunze, Manuela (2002a): An XML based Document Suite. In: Proceedings of Coling 2002, 1278 – 1282.

Rösner, Dietmar/ Kunze, Manuela (2002b): Exploiting Sublanguage and Domain Characteristics in a Bootstrapping Approach to Lexicon and Ontology Creation. In: Proceedings of OntoLex 2002 – Ontologies and Lexical Knowledge Bases at the LREC 2002, 68-73.