Eine XML-basierte Werkbank für das Document Mining*

Manuela Kunze und Dietmar Rösner

Zusammenfassung

Dieses Paper beschreibt die Ansätze, die verfolgt werden, um mit dem Tool XDOC Informationen aus Dokumenten zu extrahieren. Dabei wird XDOC um eine semantische Komponente erweitert. Die ersten Schritte in diese Richtung haben sich zunächst auf die Erkennung und Analyse einfacher Wissenstrukturen konzentriert. Linguistisch gesehen werden diese Strukturen z. B. über phrasale Muster oder durch bestimmte Nominalphrasen beschrieben. Die Darstellung der Ergebnisse der Parser, ob nun syntaktisch oder semantisch, erfolgt einheitlich in XML-Notation. Mit XML und Stylesheets ist eine flexible Möglichkeit vorhanden, relevante Informationen ohne großen Aufwand abzubilden bzw. innerhalb eines Dokumentes hervorzuheben.

12.1. Der Rahmen: DFG-Forschergruppe InfoFusion

Das hier vorgestellte Projekt ist ein Teilprojekt der DFG-Forschergruppe „Workbench für die Informationsfusion“[1]. Im Rahmen dieser Forschergruppe wird untersucht, wie mit unterschiedlichen Fusionsmethoden Informationen aus vorliegenden heterogenen Quellen zusammengebracht („fusioniert“) und Nutzern in aufbereiteter Form zur Verfügung gestellt werden können.

Im hier beschriebenen Teilprojekt wird untersucht, wie mit Techniken des document mining Informationen aus großen Beständen elektronisch verfügbarer Dokumente extrahiert und in einer geeigneten Form einem Wissensingenieur (kurz: KE) zur Verfügung gestellt werden können, so daß dieser die Ergebnisse der Extraktion ohne größeren Aufwand in eine *Wissensbasis* (KB) übernehmen kann.

Den Kern der entwickelten Werkbank für das document mining bilden die XML-basierten linguistischen Module des Systems XDOC[2]. Dieses System wurde für die Erfordernisse der gewählten Anwendungsdomäne erheblich erweitert. Dazu gehören einerseits Erweiterungen im Hinblick auf Robustheit und möglichst große Abdeckung der relevanten linguistischen Phänomene. Andere Erweiterungen betreffen Komponenten zur semantischen Analyse mit Kasusrahmen und die Integration eines XSL-Prozessors für die flexible Präsentation von Analyseergebnissen, die im XML-Format vorliegen, mit Hilfe von XSL-Stylesheets.

* Erschienen in: *Proceedings der GLDV-Frühjahrstagung 2001*, Henning Lobin (Hrsg.), Universität Gießen, 28.–30. März 2001, Seite 131–140. `http://www.uni-giessen.de/fb09/ascl/gldv2001/`

[1] `http://fusion.cs.uni-magdeburg.de`

[2] XDOC: *X*ML based tool for *doc*ument processing (Rösner, 1999, 2000)

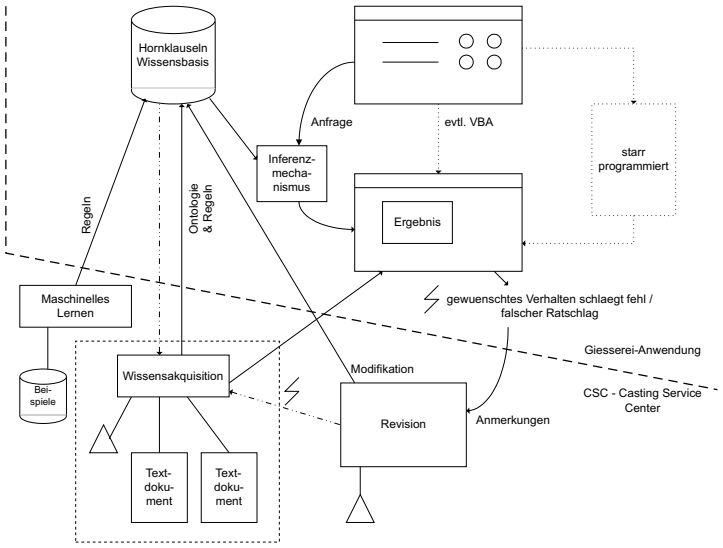

Abbildung 12.1.: Aufbau des Castings Service Centers

Abb. 12.1 zeigt das Szenario eines geplanten Systems zur Unterstützung von Gießereien (das sogenannte *Casting Service Center*, CSC) und die Einordnung der Akquisition von Wissen aus Dokumenten innerhalb dieses Rahmens.

Innerhalb der Gießereitechnologie besteht, wie in den meisten anderen technologischen Bereichen, das Problem, mit einer großen Wissensmenge zu arbeiten. Der KE muß mit einer Vielfalt von Wissensarten umgehen können. Innerhalb des hier beschriebenen Projektes haben erste Untersuchungen gezeigt, daß folgende Arten von Wissen verarbeitbar sein sollten:

- Beschreibungen von Kategorien und Instanzen,
- Produktionsregeln („Wenn-Dann“-Regeln),
- Beschreibungen von zeitlichen Abläufen bzw. Situationsbeschreibungen,
- Constraints (bedingt z. B durch die Parameter der genutzten Maschinen),
- unscharfes Wissen.

Die Fragestellung unseres Projektes ist: inwieweit ist es möglich, die oben genannten Wissensarten in unserem Korpus zu extrahieren und wie kann eine effektive Unterstützung des KE realisiert werden.

Die in der Abb. 12.1 dargestellte Wissensgewinnung durch Maschinelles Lernen wird in einer parallelen kooperierenden Arbeitsgruppe untersucht und wird an dieser Stelle nicht weiter ausgeführt.

12.2. Charakterisierung des Korpus

Für verschiedene Teilprojekte der Forschergruppe wurde das Gießereiwesen als gemeinsame Domäne gewählt, da hier sehr heterogene Informationsbestände vorliegen.

Als Material für die Extraktion von Gießereiwissen stehen uns als Textbestand (Korpus) einerseits Kapitel aus Fachbüchern über Gießereitechnik und andererseits verbal formulierte Arbeitsanleitungen bzw. Aufstellungen zu beachtender Regeln zur Verfügung. Die Dokumente lagen im Original größtenteils in Word-Dokumenten vor. Bei der Konvertierung der Dokumente in ASCII-Files mußte ein teilweiser Verlust von (im Layout) enthaltener Strukturinfo sowie von Bildern, Tabellen und anderer nicht-textueller Objekte in Kauf genommen werden. Die Spezifika der einzelnen Korpusbeständen werden im folgenden vorgestellt.

12.2.1. Fachbuchtexte

Bei den Fachbuchtexten handelt es sich um grundlegende Beschreibungen der Domäne, die zur Einarbeitung in das Fachgebiet gedacht und geeignet sind. Es werden zunächst die allgemeinen Konzepte (begriffliches Wissen) beschrieben, und darauf aufbauend konkrete Instanzen der zuvor beschriebenen Konzepte erläutert.

In diesen Texten werden vielfältige Textarten (Fließtexte, Tabellen, Formeln, Fußnoten etc.) zur Beschreibung der Domäne genutzt. Aus linguistischer Sicht ergeben sich daraus interessante Aufgaben. Neben dem Auflösen der üblichen Referenzen innerhalb eines Textes, liegen hier als Metaobjekte Verknüpfungen zu anderen Objekten vor, die durch den Export der ursprünglichen Daten zum Teil nicht mehr im Dokument vorhanden sind bzw. interpretierbar sind, z.B.: *Wie Bild 2.1 zeigt . . .* oder *der Mittelwert nach Gleichung (2) . . . .*

Bei der Nutzung dieser Dokumente durch den KE, sollte es auch möglich sein, diese Metaobjekte auszuwerten. Da unser System als interaktives System geplant ist, könnte man die Referenzen auf diese Metaobjekte als einen Hyperlink markieren. Während der Interaktion mit dem Benutzer kann dieser über diesen Link auf das Metaobjekt zugreifen und gegebenenfalls die fehlenden Informationen, die innerhalb des Metaobjektes beschrieben sind, manuell eintragen. Mit diesen Links wird es dann z. B. auch möglich auf innerhalb des Fließtextes referenzierte Literatur, soweit sie elektronisch verfügbar ist, zuzugreifen.

Die Diskursstruktur spielt eine wichtige Rolle bei der Interpretation der Texte. So werden z. B. häufig Klammerausdrücke genutzt, aber für sehr unterschiedliche Funktionen, zum Beispiel als erklärende Komponente oder als Literaturverweis. Beispiele sind:

- als Instanz eines Konzept, wenn z. B. ein Beispiel angegeben wird: *in nicht metallische (vorwiegend aus Quarzsand) oder metallische Formen . . .*
- Literaturverweis: *In Anlehnung an die Literatur (Ambos, E.; ISBN 3-342-00379-0) . . .*
- Numerierungen von Objekten sowie die Referenzierung, in diesem Fall Formeln *nach Gleichung (2) . . .*
- Beziehungen zwischen Konzepten: *. . . mit denen auf den Formwerkstoff (bei verlorenen Formen) oder auf das flüssige Metall (bei Dauerformen) . . .*
- Auflösung von Abkürzungen: *Gefüge von GGL (Gußeisen mit Lamellengraphit) . . .*
- erklärende Funktionen (Spezifizierung): *in der gewünschten Wanddicke grau, weiß oder meliert (Übergangszone zwischen Weiß- und Grauerstarrung) . . .*

12.2.2. Arbeitsanleitungen

Diese Texte setzen schon Vorkenntnisse aus dem Bereich voraus, so daß sie weniger zur Einarbeitung in die Domäne geeignet sind. Hier werden sehr viele Termini aus der Subsprache verwendet und die Struktur der Texte ist eine ganz andere als bei den Fachbuchtexten. Die Anleitungen beginnen meistens mit einer Regel oder einem Leitsatz, gefolgt von einer Liste von Begriffen, die als Schlüsselbegriffe den Fokus der Regel genauer beschreiben. Danach folgt eine Erläuterung bzw. Begründung der Regel. Nach diesen Begründungen können Lösungsbeschreibungen folgen, die Anleitungen für solche Fälle darstellen, in denen der zuvor genannte Leitsatz nicht oder nicht vollständig eingehalten werden konnte. Die Regelmenge ist nach Anwendungsschwerpunkten gruppiert. Nachfolgend ist ein Beispiel für eine Regel aus diesem Korpus aufgeführt.

Vermeide das Kreuzen von Kernen!

Dauerform, Kerne, Kernkreuzung, modell- und formenbaugerecht

Das Kreuzen von Kernen birgt die Gefahr der Kollision von Kernzügen und fördert das Auftreten von Grat, der mit zusätzlichem Aufwand zu entfernen ist. Erfordert die Funktion eines Gußstücks das Aufeinanderstoßen von Bohrungen oder anderer Öffnungen, dann ist durch zweckmäige Gestaltung des Teils das Kreuzen von Kernen zu vermeiden.

12.3. Aktuelle Arbeiten

Eine zentrale Frage für das Projekt ist, mit welchen sprachlichen Strukturen in den Texten des Beispielkorpus solche Beziehungen ausgedrückt werden, die für eine Modellierung in einem Formalismus zur Wissenrepräsentation besonders relevant sind. Dazu gehören u. a. Begriffsdefinitionen, Teil-Mengen-Beziehungen und Regeln. Im folgenden werden einige markante Beispiele kurz beschrieben. Da in unseren Arbeiten mit Ontologien gearbeitet wird, soll hier eine kurze Begriffsklärung folgen.

Ontologie

Die bekannteste Definition von „Ontologie“ stammt von Gruber (1993): „An Ontology is an explicit specification of a conceptualization.“

Ursprünglich stammt der Begriff Ontologie aus der Philosophie (Lehre des Seins). Innerhalb der Informatik wird mit einer Ontologie die Darstellung und Formalisierung von Wissen beschrieben. Durch eine Ontologie werden Teile der realen Welt über abstrakte Konzepte und Rollen (Relationen) zwischen den Konzepten beschrieben. Eine Ontologie entspricht einem abstrakten Modell eines Weltauschnittes, wobei der Umfang und der Bereich der Modellierung abhängig vom Verwendungszweck ist.

Für unsere Arbeit wurden bei der Modellierung einer Ontologie technischer Zusammenhänge vorwiegend die Konzepte zur Beschreibung von Prozeßen untersucht. Ein Beispiel für die Einbindung dieser Ontologie bei der Extraktion von Informationen ist in 12.3.3 zu finden.

12.3.1. Phrasale Muster

Derzeit wird mit Hilfe der entwickelten Werkzeuge – insbesondere dem robusten partiellen Parser von XDOC – in den Texten nach phrasalen Mustern gesucht, die „interessante Beziehungen“ im

obigen Sinne ausdrücken, und es wird überprüft, als wie treffsicher sich die aus den Vorkommen abstrahierten Generalisierungen in der Domäne erweisen.

Beispiele:

> *Als formlose Stoffe gelten Gase, Flüssigkeiten, Pulver, Fasern, Späne, Granulate, Lösungen, Schmelzen u. ä.*
>
> *. . . die Ausgangsmaterialien (Roheisen, Schrott, Ferrolegierungen u. ae.) . . .*

führt zum generalisierten Muster

> *Als <Kollektivbegriff > gelten <Enumeration von Kollektivbegriffen>.*

bzw.

> *<Kollektivbegriff >(<Enumeration von Kollektivbegriffen>)*

Letztgenanntes Muster ist relativ häufig im Korpus zu finden. Dem Wissensingenieur können die Vorkommen dieser Muster in den Dokumenten dann Kandidaten für die Definition von Begriffen als Überdeckungen anderer Begriffe liefern.

Ein anderes phrasales Muster läßt sich durch folgenden Beispielsatz beschreiben:

> *Als Handformen wird die Herstellung einer Sandform ohne Benutzung einer Formmaschine bezeichnet.*

Hierbei wird kein Kollektivbegriff beschrieben, sondern die Definition eines Konzeptes vorgenommen. Formal beschrieben, sieht das Muster dann wie folgt aus:

> *Als <Konzept> wird <Definition> bezeichnet.*

Zu diesen Beschreibungen gehört auch folgende Formulierung:

> *<Konzept> wird wie folgt definiert: <Definition>.*

12.3.2. Normierte Bezeichner

Jede Domäne arbeitet mit gebietsspezifischen sogenannten normierten Bezeichnern. In der Medizin sind dies z. B. Kennnummern für Krankheiten (ICD-Code) aber auch für Enzyme oder Gennotationen. Diese Normierungen werden innerhalb der Domäne häufig genutzt und deren Aufbau kann meist über Grammatikregeln beschrieben werden. In dem vorliegendem Korpus können z. B. folgende Arten normierter Bezeichner gefunden werden:

Produktbezeichner: Gußstück EN 1982 – CC333G – GS – XXXX

Werkstoffe für Legierungen: EN-GJL-150

Gußwerkstoffe: EN-GJMW-350-4

Legierungstypen: G-Al Mg

Diese Token können mit einem gebietsspezifischen Erkenner für Bezeichner von Gießereiprodukten detektiert werden. Als Beispiel soll hier die Struktur des Produktbezeichners vorgestellt werden.

Der Erkenner liefert hier eine Struktur vom Typ `<PRODUCT>`. Der Aufbau eines Produktbezeichners setzt sich aus mehreren Tags zusammen und läßt sich wie folgt beschreiben:

`<N> <NORM> - <Mat-ID> - <Methode> - <Modellnr>.`

Der *Nomen*-Tag am Anfang der Struktur entspricht einer Benennung des Produktes (z. B. Blockmetall oder Gußstück). Die weiteren Tags lassen sich wie folgt beschreiben:

Norm: Norm, in der die Materialkennung beschrieben wurde.

Mat-ID: Materialbezeichner in Form einer Werkstoffkennnummer (CB333G) oder chemischen Formeln (CuAl10Fe5Ni5-B).

Methode: Bezeichnung des Gießverfahrens (mit GS wird z.B. der Sandguß beschrieben). Diese Kennzeichnungen wurden in der ISO 1190-1 festgelegt.

Modellnr: die Modell-, Form- oder Zeichnungsnummer des Gußstückes.

Durch eine Erweiterung der Regeln des Parsers (Beschreibung des Konstrukts *PRODUCT*) und ein entsprechendes Lexikon (Aufnahme der gebietsspezifischen Bezeichner wie z. B. denjenigen für Gießverfahren) können diese Konstrukte (z. B. Mat-ID) erkannt und interpretiert werden. Dadurch erhalten wir erste semantische Informationen bzw. Tags, die semantischer Art sind.

Bsp. 1
```
<PRODUCT Method="Sandguss" Material="CC333G">
    <N>Gussstueck</N>
    <NORM>
      <N>EN</N>
      <NR>1982</NR>
    </NORM>
    <IP>-</IP>
    <MAT-ID>CC333G</MAT-ID>
    <IP>-</IP>
    <METHODE>GS</METHODE>
    <IP>-</IP>
    <MODELLNR>XXXX</MODELLNR>
  </PRODUCT>
```

Es ist noch zu prüfen, ob durch die Übernahme der Features der eingebetteten Strukturen in den Tag *PRODUCT* ein Vorteil bei der Weiterverarbeitung zu erzielen ist. In der Grammatik von XDOC können *PRODUCT*-Strukturen dann als Nomen weiterverarbeitet werden und den Kern von Nominalphrasen bilden.

12.3.3. Nominalisierte Verben

Da innerhalb des Korpus relativ häufig nominalisierte Verben genutzt wurden, haben wir uns in der hier beschriebenen Arbeit zunächst darauf konzentriert, die Vorkommen dieser Wortart genauer zu analysieren.

Im Korpus treten häufig Nominalphrasen auf, die meist Attachments in Form von Präpositionalphrasen und Nominalphrasen im Genitiv enthalten. Die richtige Zuordnung der Phrasen zum Nomen realisieren wir mit Hilfe der Ontologie und der Kasusrahmen. Neben dem Lexikon mit den syntaktischen Merkmalen wurde bzgl. der Nomen ein Lexikon zur Beschreibung der semantischen Merkmale erstellt. In diesem Lexikon werden zu einem Nomen die möglichen semantischen Relationen (Kasusrollen) und ihre syntaktische Form (Kasusform) beschrieben. Dadurch sollte es möglich sein, das Problem des PP-Attachments zu lösen. Bei einer nicht eindeutigen Zuweisung der PP zum Nomen bzw. zum Verb wäre es denkbar, einen statistischen Ansatz, wie er z. B. in Volk (2000) beschrieben wurde, hinzu zu ziehen oder über Benutzerinteraktion das Problem zu lösen. Die im Lexikon beschriebenen Kasusrahmen der Nomen können systematisch vom dazu gehörigen Verb abgeleitet werden.

Ein kleiner Ausschnitt aus der zugrunde liegenden Ontologie ist in Abb. 12.2 zu sehen.

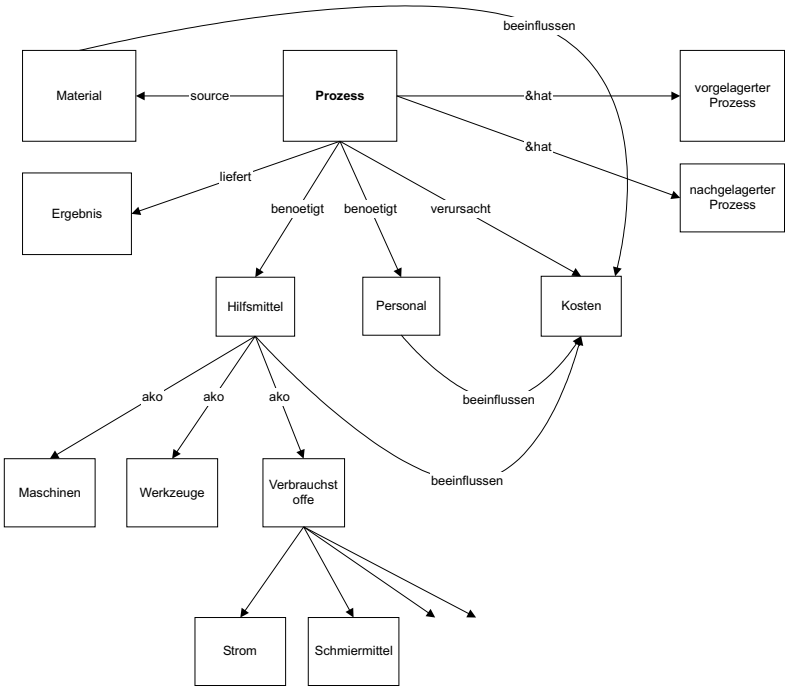

Abbildung 12.2.: Teil der Domänenmodellierung

In dem hier beschriebenen Ansatz werden linguistische Analyse-Methoden und semantische Informationen über Kasusrollen und Relationen der Objekte der realen Welt miteinander kombiniert. Anhand des Beispielsatzes

„*Nach DIN 8580 ist Urformen das Fertigen eines festen Koerpers aus formlosem Stoff durch Schaffen des Zusammenhalts.*“

soll der Ansatz kurz vorgestellt werden. Die in der Ontologie beschriebenen Relationen wurden in

die Lexika übernommen. Bei Nominalisierungen kann man davon ausgehen, daß ein Konzept aus der Kategorie „Prozeß“ oder aus der Kategorie „Ergebnis“ (z. B. *das Schaffen Goethes*) beschrieben werden. Im Lexikon werden durch die beschriebenen Valenzforderungen die Relationen aus der Ontologie übernommen. Die Beschreibung des Konzeptes *Fertigen* innerhalb des semantischen Lexikons sieht z. B. wie folgt aus:

```
XDOC(9): (Print-sem "fertigen")
Zuordnung fuer "fertigen"
Beschreibung: "Schaffung von etwas"
Rollenzuordnung: "Prozess"

Valenzforderungen
Beschreibung: "instrument"
              sem. Ford.: "Prozess"
              syn. Ford.: "P(akk,fak,durch)"
Beschreibung: "source"
              sem. Ford.: "Material"
              syn. Ford.: "P(dat,fak,aus)"
Beschreibung: "result"
              sem. Ford.: "Objekt"
              syn. Ford.: "N(gen, fak) P(akk,fak,von)"

Format: "fertigen(_result_,_source_,_instrument_)"
```

Innerhalb des Lexikons werden die semantischen Forderungen (Prozess, Material, Objekt) der Attachments mit ihren syntaktischen Ausprägungen beschrieben. Für die Kasusrolle *result* gilt z. B. die semantische Forderung, daß der Rollenfüller eine Instanz des Konzeptes *Objekt* ist. Syntakisch wird für diese Rolle eine Nominalphrase im Genitiv oder eine Präpositionalphrase im Akkusativ mit der Präposition *von* gefordert.

Die Prüfung der semantischen Forderungen setzt voraus, daß jedes im Korpus auftretende Nomen eine Zuordnung zu den Kategorien innerhalb des semantischen Lexikon besitzt. Bei einem unvollständigem Lexikon wird es dem Benutzer ermöglicht, über eine Interaktionsschnittstelle das Lexikon manuell zu ergänzen. Die zur Zeit erfaßten Einträge im semantischen Lexikon wurden anhand der Beschreibungen in der initialen Ontologie und verschiedenen Lexika (z. B. Helbig und Schenkel, 1991, Sommerfeldt und Schreiber, 1977, Germanet, vgl. `http://www.sfs.nphil.uni-tuebingen.de/lsd/`), und durch Auswertung der im Korpus vorkommenden Merkmale der Begriffe erstellt.

Das Ergebnis der Zuordnung der Ergänzungen zum Wort *Fertigen* zu den definierten semantischen Rollen ist in Abb. 12.3 zu sehen.

Die Auszeichnung der semantischen Informationen des Definitionsteils des Beispielsatzes ergibt die XML-Struktur in Beispiel 2.

Bsp. 2
```
<PROZESS>
        <NAME>Fertigen</NAME>
        <SOURCE>aus formlosem Stoff</SOURCE>
        <RESULT>von festen Koerpern</RESULT>
        <INSTRUMENT>durch
```

```
      <PROZESS>
        <NAME>Schaffen</NAME>
        <RESULT>des Zusammenhalts</RESULT>
      </PROZESS>
    </INSTRUMENT>
  </PROZESS>
```

Die Bezeichner der Tags entsprechen den in der Ontologie beschriebenen Kategorien. Es werden den verschiedenen Phrasen die Rollen der Kasusrahmenanalyse zugeordnet. Durch diese Art der Strukturierung wird nicht nur die Präsentation, sondern auch die Weiterverarbeitung für den KE erleichtert. Durch den Tag *Prozess* werden die zusammengehörigen Konzepte und Relationen als eine Einheit betrachtet und analysiert. Die Übernahme der gefundenen Konzepte in die Wissensbasis erfolgt nicht automatisch.

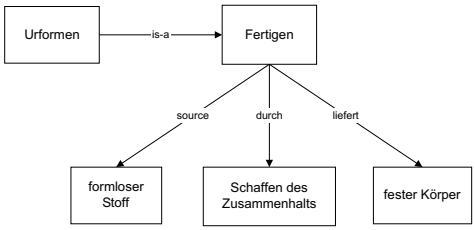

Abbildung 12.3.: Interpretation des Satzes *Nach Din 8580 ist Urformen Fertigen ...*

In die KB werden nur die Relationen übernommen, die auch tatsächlich im Text beschrieben worden sind, d. h. es kann auch der Fall passieren, daß die abgebildeten semantischen Informationen nicht korrekt bzw. nicht vollständig sind. In diesen Fällen ist die Interaktion mit dem KE notwendig. Der KE prüft die Einträge auf Vollständigkeit und Richtigkeit. Gegebenenfalls kann er dann manuell die Daten korrigieren bzw. vervollständigen.

12.4. Zusammenfassung

Neben den gebietsunabhängien Komponenten (Tagger, Chunker, Parser, ...) des XDOC-Systems benutzen wir zur Analyse der Texte aus dem Korpus *Gießertechnik* gebietsspezifische Erkenner und semantische Lexika, in denen die Verknüpfung zwischen syntaktischen und semantischen Bedingungen die auftretenden Konzepte und Begriffe realisiert wird. Der Aufbau dieser Lexika wird durch die Übernahme von Relationen und Konzepten der initialen Ontologie realisiert. Über Kasusrahmen werden den übernommenen Relationen syntaktische Merkmale im Lexikon zugeordnet.

Alle Komponenten von XDOC (Rösner, 2000) liefern ihre Ergebnisse als XML-Strukturen. Diese Design-Entscheidung war bewußt getroffen worden. Für die Aufgaben der Pattern-Extraktion hat sie sich bereits in vielfacher Weise „ausgezahlt“: mit XSL-Stylesheets[3] wird festgelegt, wie die Ergebnisse der Analyse von XDOC dem Betrachter (in einem beliebigen Webbrowser) dargeboten werden sollen, d. h. welche Teilstrukturen interessieren und auf welche Weise sie daher

[3] `http://www.w3.org/Style/XSL`

hervorzuheben sind. Mit dem XT-Paket von James Clark[4] wird die Ausgabe nach den Vorgaben des Stylesheet dann ohne zusätzlichen Programmieraufwand bewerkstelligt.

Eine offene Frage ist, ob für jede linguistische Struktur nur eine Ergebnisstruktur produziert werden soll, in der Ergebnisse der syntaktischen und der semantischen Analyse zusammengefaßt sind oder ob es vorteilhafter ist, getrennte Ergebnisstrukturen zu verwenden.

Ein ähnlicher Ansatz zur Markierung von syntaktischer und semantischer Informationen findet sich zum Beispiel in GATE[5]. Dieses Tool verwenden wir zur Zeit innerhalb unserer Projektarbeit für eine äquivalente Aufgabenstellung mit englischen Texten aus dem medizinischen Bereich.

Literaturverzeichnis

Gruber, T. R. (1993): "A Translation Approach to Portable Ontology Specifications". *Knowledge Acquisition* 5 (2): S. 199–220.

Helbig, G. und Schenkel, W. (1991): *Wörterbuch zur Valenz und Distribution deutscher Verben*. Bibliographisches Institut Leipzig, 8. Auflage.

Rösner, D. (1999): "XDOC – XML-basierte Werkzeuge für multilinguale Korpora". In: *Multilinguale Corpora – Codierung, Strukturierung, Analyse; 11. Jahrestagung der GLDV*. Prag: Enigma, S. 332–341.

Rösner, D. (2000): "Combining Robust Parsing and Lexical Acquisition in the XDOC System". In: *KONVENS 2000: 5. Konferenz zur Verarbeitung natürlicher Sprache*. Ilmenau, S. 75–80.

Sommerfeldt, K. E. und Schreiber, H. (1977): *Wörterbuch zur Valenz und Distribution der Substantive*. Bibliographisches Institut Leipzig.

Volk, Martin (2000): "Scaling up. Using the WWW to resolve PP Attachment Ambiguities". In: *KONVENS 2000: 5. Konferenz zur Verarbeitung natürlicher Sprache*. Ilmenau, S. 151–156.

[4] `http://www.jclark.com`
[5] `http://gate.ac.uk`