Generierung von Linkangeboten zur Rekonstruktion terminologiebedingter Wissensvoraussetzungen

Michael BEISSWENGER
Universität Dortmund, Institut für deutsche Sprache und Literatur, D-44221 Dortmund
beisswenger@hytex.info

Eva Anna LENZ
Universität Dortmund, Institut für deutsche Sprache und Literatur, D-44221 Dortmund
lenz@hytex.info

Angelika STORRER
Universität Dortmund, Institut für deutsche Sprache und Literatur, D-44221 Dortmund
storrer@hytex.info

Abstract

Dieser Beitrag skizziert Strategien zur (semi-)automatischen Annotation von definitorischen Textsegmenten und Termverwendungsinstanzen auf der Grundlage grammatisch annotierter Korpora. Ziel unserer Überlegungen ist es, bei der selektiven Rezeption von Fachtexten in einer Hypertextumgebung die je spezifischen Wissensvoraussetzungen, die der Verwendung von Fachtermini unterliegen und die für das Textverständnis eine entscheidende Rolle spielen, über automatisch generierte Linkangebote rekonstruierbar zu machen.

1 Projektrahmen und Szenario

Im Projekt »Hypertextualisierung auf textgrammatischer Grundlage« (*HyTex*, vgl. http://www.hytex.info) geht es um die (semi-)automatische Generierung eines hypertextuell organisierten Präsentationsformats für ein Korpus mit sequenziell organisierten Fachtexten zur Domäne Texttechnologie. Bei der Erzeugung dieser als Hypertextsichten bezeichneten Präsentationsformate sollen den Nutzern beim Browsen genau diejenigen Wissensvoraussetzungen angeboten werden, die jeweils zum Verständnis des aktuell rezipierten Inhalts benötigt werden. In diesem Kontext werden automatische Strategien zur Segmentierung und zum Linking entwickelt und am Korpus getestet. Die Strategien zur Generierung von Hypertextsichten operieren über Repräsentationen von Wissen auf dreierlei Ebenen, die in der Veranschaulichung der Projektarchitektur in Abb. 1 (von unten nach oben) dargestellt sind:

– Das in den Dokumenten sprachlich manifeste Wissen über die Vernetztheit der Inhalte wird repräsentiert als textgrammatisches und linguistisches Markup, das auf eine grammatische Vorannotation aufsetzt, die vom DEREKO-Projekt[1] bereit gestellt wird. Die Vorannotation (Part-of-Speech-Tagging, Chunk-Annotation) wird erweitert um Annotationen solcher textgrammatischer Einheiten, die für die Generierung kohäsiv geschlossener Hypertext-Module relevant sind (insbesondere Koreferenz und Konnexion).

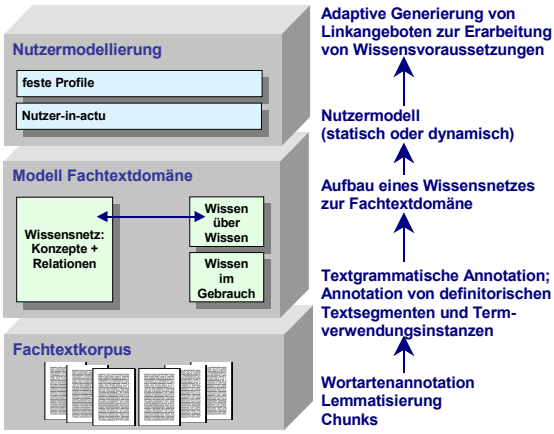


Abb. 1: Die *HyTex*-Projektarchitektur.

– Das Wissen über die Relationen zwischen zentralen Konzepten und Termini der Fachtextdomäne wird mit XML Topic Maps (XTM, [PM01]) modelliert, einem auf XML-Syntax basierenden Standard zur Beschreibung von Metadaten, die in Form eines Netzwerks organisiert sind (vgl. [LBS02]).
– Annahmen über das Vorwissen bestimmter Nutzertypen werden zunächst als (statische) Nutzerprofile modelliert, in einer späteren

[1] Zu DEREKO (»Deutsches Referenzkorpus«) siehe http://www.sfs.nphil.uni-tuebingen.de/dereko/.

Projektphase dann auch in Form von Nutzungsprotokollen, aus denen sich dynamisch die Wissensvoraussetzungen erschließen lassen, die ein Nutzer auf seinem individuell gewählten Leseweg bereits erworben hat.

Unter Rückgriff auf Wissen aus diesen drei Ebenen werden Strategien implementiert, um (a) Hypertext-Sichten auf die Korpustexte zu generieren, deren einzelne Module kohäsiv geschlossen sind, und (b) Wissensvoraussetzungen, die der Verwendung von (Fach-)Termini in den Dokumenten zu Grunde liegen, über Linkangebote zu den entsprechenden Definitionen für einen selektiv zugreifenden Benutzer rekonstruierbar zu machen.

Ad a) Die textgrammatische Annotation erfolgt teilautomatisch. Die Erarbeitung des hierzu notwendigen Tagsets ist Teil der Projektarbeit.

Ad b) Die Generierung von Linkangeboten zur Rekonstruktion von terminologiebedingten Wissensvoraussetzungen erfolgt auf der Grundlage einer ebenfalls teilautomatischen Annotation sowohl von Termverwendungsinstanzen als auch von Textsegmenten, in deren Rahmen Termini definitorisch eingeführt werden.

In diesem Beitrag konzentrieren wir uns ausschließlich auf Strategien zur Annotation von definitorischen Textsegmenten und Termverwendungsinstanzen. Ziel ist es aufzuzeigen, wie in einer Hypertextumgebung für die Rezeption von Fachtexten solchen Kohärenzproblemen vorgebeugt werden kann, die sich durch die selektive Lektüre ergeben (vgl. [Sto02]). Diese Probleme sind darin begründet, dass ein Rezipient in Bezug auf einen für ihn nicht oder nicht exakt semantisierbaren sprachlichen Ausdruck zu entscheiden hat,

1. ob es sich bei dem betreffenden Ausdruck um einen Terminus handelt oder nicht;
2. wenn es sich um einen Terminus handelt:
   2.1 ob dieser vom Autor relativ zu einem ganz bestimmten (im Vortext explizit oder implizit eingeführten) Konzept verwendet wird und eine entsprechende Definition daher durch »Zurückblättern« zu suchen ist, oder
   2.2 ob dieser vom Autor relativ zu einem in der Fachsprache etablierten Konzept verwendet wird und eine entsprechende Definition daher nicht im Vortext, sondern beispielsweise in einem einschlägigen Fachwörterbuch zu suchen ist.

Eine Hypertextumgebung kann die selektive Textrezeption dahingehend unterstützen, dass sie – bei entsprechender Qualität des zu Grunde liegenden Korpus – Linkangebote bereitstellt, die (i) die angeführten Entscheidungsnotwendigkeiten hinfällig machen und (ii) auf diejenigen Textstellen konnektiert sind, aus welchen sich die für die jeweilige Verwendung eines Terminus relevanten semantischen Informationen erschließen lassen (also im Fall 2.1 auf entsprechende definitorische Passagen im Vortext, im Fall 2.2 auf entsprechende Einträge in einem Fachwörterbuch).

Für die Implementierung unserer Architektur nutzen wir XML als Austauschformat für alle Komponenten: für das Markup des Textkorpus, für die Topic Map (XTM-Syntax) und für die Benutzermodellierung. Zur Erzeugung der späteren Hypertextsichten aus der textgrammatischen Annotation und der Topic Map benutzen wir XSLT; zu Einzelheiten siehe [LS02].

2 Grundlagen einer teilautomatischen Annotation definitorischer Textsegmente

Um bei der Generierung von Hypertext-Sichten automatisch Links zu den jeweils relevanten Definitionen legen zu können, unterscheiden wir in Bezug auf die Verwendung von Termini in Fachtexten drei verschiedene Arten von Wissensvoraussetzungen:

- Eine *intratextuelle Wissensvoraussetzung* liegt vor, wenn ein Ausdruck als Terminus im Sinne einer Definition verwendet wird, die der Autor im Vortext explizit eingeführt hat.
- Eine *extratextuelle Wissensvoraussetzung* liegt vor, wenn der Autor einen Terminus im Sinne der Definition eines anderen Autors (in einem anderen Dokument) verwendet.
- Eine *domänenspezifische Wissensvoraussetzung* liegt vor, wenn ein Terminus ohne genaue Angabe einer Definitionsstelle im Sinne einer in der betreffenden Fachsprache eingespielten Festlegung verwendet wird.

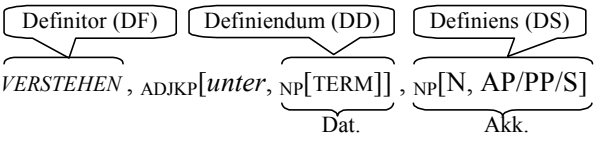
Um zu ermitteln, (a) welche der drei Arten von Wissensvoraussetzungen der Verwendung eines Terminus unterliegt, und (b) an welcher Stelle des betreffenden Dokuments im Falle einer intratextuellen Wissensvoraussetzung dasjenige

definitorische Textsegment zu finden ist, welches als Zielanker eines entsprechenden Linkangebots in Frage kommt, müssen nicht nur die Termverwendungsinstanzen, sondern auch sämtliche definitorischen Textsegmente des *HyTex*-Korpus annotiert und typisiert sein.

Die Annotation der Termverwendungsinstanzen erfolgt automatisch auf der Grundlage einer manuell erarbeiteten Termkandidatenliste zur Fachtextdomäne Texttechnologie. Die Identifizierung und Annotation der definitorischen Textsegmente im Korpus erfolgt teilautomatisch und auf der Grundlage sowohl einer funktionalen Typologie von Definitionen als auch einer Beschreibung grammatischer Strukturmuster für definitorische Textsegmente mit den im Deutschen als Definitoren verwendeten Prädikatoren (wie z.B. *bezeichnen (als)*, *verstehen (unter)* mit Adjunktorphrase oder Kopula *sein* in Prädikativkonstruktionen mit definiensartiger Struktur der prädikativen Ergänzung und generisch referierendem Subjekt).

Eine Typologie von Definitionen ist deshalb notwendig, um in Fällen zweier oder mehrerer konkurrierender Definitionen eines Terminus in ein- und demselben Fachtext ermitteln zu können, welcher davon unter handlungssemantischem Aspekt die höchste Priorität zugesprochen werden kann. Grammatische Strukturmuster sind zum einen notwendig, um solche Fälle auszuschließen, in welchen ein Prädikator (z.B. *bezeichnen (als)*), der in bestimmten Verwendungen als Definitor fungieren kann, mit nicht-definitorischer Intention gebraucht wird (beispielsweise in Sätzen wie »Die Bedeutungskonzeption von N.N. muss als hochgradig problematisch bezeichnet werden«). Zum anderen lässt sich anhand der grammatischen und syntaktischen Eigenschaften eines definitorisch und in einer bestimmten flexivischen Ausprägung verwendeten Prädikators automatisch identifizieren, in welcher syntaktischen Position des betreffenden Textsegments diejenigen Phrasen zu lokalisieren sind, die im Rahmen der Definition als Definiendum bzw. als Definiens fungieren.

Beispiel: Allgemeines Muster für definitorische Textsegmente mit dem Prädikator *verstehen* + Adjunktorphrase (vereinfacht):



»ADJKP[Unter **DD**(NP)[einem TERM[Link]]] **DF**[verstehe] ich **DS**(NP)[eine AP[computerverwaltete] N[Zuordnung] PP[zwischen Ankern]].«

3 Pragmatische Fundierung

Die für die teilautomatische Annotation von definitorischen Textsegmenten zu Grunde gelegte Typologie von Definitionen geht aus von einer handlungssemantischen Konzeption der Versprachlichung definitorischer Zuordnungen in der Fachkommunikation: Definitorische Äußerungen sind kommunikative Initiativen, die auf einem *definitorischen Sachverhaltsentwurf* basieren, der zur Erreichung eines bestimmten *kommunikativen Zwecks* und unter Rückgriff auf ein für diesen Zweck verfügbares *Handlungsmuster* hervorgebracht wird, für den ein be-

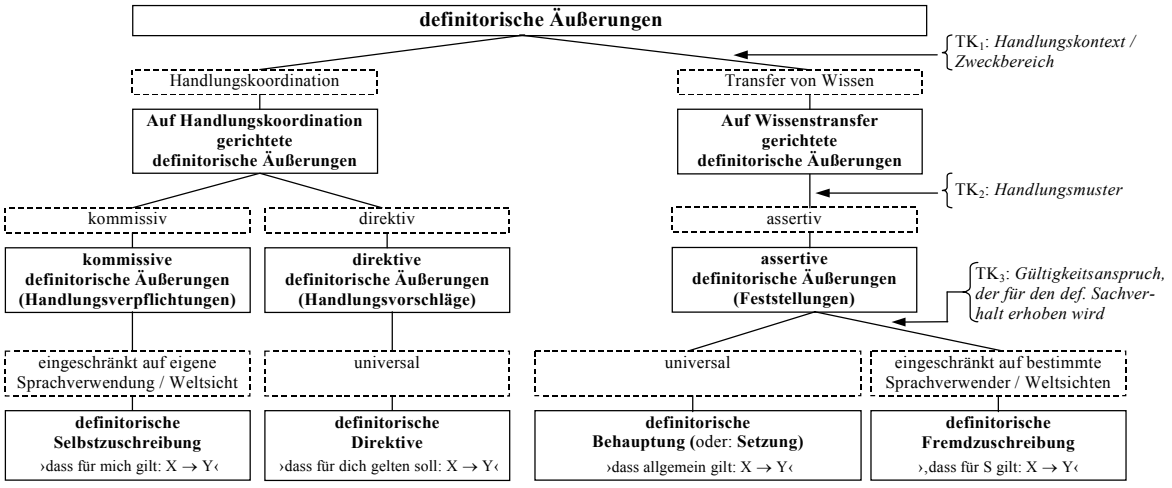


Abb. 2: Ansatz zu einer handlungssemantisch motivierten Typologie definitorischer Äußerungen.

stimmter Gültigkeitsanspruch erhoben wird, und dessen Versprachlichung (relativ zum gewählten Handlungsmuster) an einem bestimmten *Versprachlichungsmuster* orientiert ist (vgl. hierzu und zum folgenden ausführlich [Bei02]).

Die *Handlungskontexte*, welchen die sprachliche Etablierung definitorischer Sachverhaltsentwürfe unterworfen sein kann, lassen sich auf oberster Ebene unterscheiden in (i) Kontexte, die auf Handlungskoordination und (ii) Kontexte, die auf einen Transfer von Wissen gerichtet sind (vgl. Abb. 2). An diese Handlungskontexte lassen sich pragmatisch bestimmte Typen definitorischer Äußerungen anschließen, denen je spezifische *Handlungsmuster* zugrunde liegen: Im Bereich der Handlungskoordination können Definitionen entweder in Form von *Kommissiven* (oder: *Selbstzuschreibungen*) vorgebracht werden, die dazu dienen, Festlegungen in bezug auf das eigene (zukünftige) Sprachhandeln zu treffen, oder in Form von *Direktiven*, die auf das Sprachhandeln der Adressaten und damit auf eine Veränderung des Sprachgebrauchs in der betreffenden Fachdomäne gerichtet sind. Im Bereich des Wissenstransfers haben definitorische Äußerungen immer die Form von *Assertiven*; diese können – je nach dem Gültigkeitsanspruch, welcher propositional für den definitorischen Sachverhalt erhoben wird – subdifferenziert werden in *definitorische Behauptungen* (oder: *Setzungen*) mit universalem Gültigkeitsanspruch und in *definitorische Fremdzuschreibungen*, bei welchem unter expliziter Angabe einer Sprachverwendergruppe oder einer Weltsicht die Gültigkeit des definitorischen Sachverhalts auf bestimmte Kontexte eingeschränkt wird.

Der nachfolgende Codeausschnitt gibt ein (vereinfachtes) Beispiel für die XML-Annotation eines definitorischen Textsegments vom Typ »Selbstzuschreibung«, in dem zwei Termini erwähnt werden, nämlich einer (*Link*) als Definiendum und ein weiterer (*Anker*) innerhalb des Definiens:

```
<DEF term="Link"
   defType="Selbstzuschreibung"
   conceptualizationNumber="1">
   Unter
   <definiendum>
      einem
      <TERM baseform="Link">
      Link</TERM>
   </definiendum>
   verstehe ich
   <definiens>
      eine computerverwaltete Zuordnung
      zwischen
      <TERM baseform="Anker"
      usageType="intra-textual"
      conceptualizationNumberRef="3">
      Ankern</TERM>
   </definiens>
</DEF>
```

4 Terminologiesensitives Linking

Die (semi-)automatische Identifizierung und Annotation definitorischer Textsegmente wird im *HyTex*-Projekt dazu genutzt, Verwendungsinstanzen von Termini auf der Interaktionsebene Linkangebote zuzuordnen, die jeweils zu derjenigen Definition führen, anhand welcher das für die korrekte Semantisierung des terminologischen Ausdrucks benötigte Konzept- und Bedeutungswissen erschlossen werden kann. Die hierbei verfolgte Strategie basiert zwar zunächst auf einer Suche nach spezifischen grammatischen Strukturmustern; aufgrund der in einem zweiten Schritt verfolgten Typisierung definitorischer Textsegmente nach Sprachhandlungsmustern geht sie aber über eine »bloße« Mustererkennung hinaus. Dies unterscheidet den skizzierten Ansatz von dem für das Englische entwickelten und in [KM01, MK02] beschriebenen Verfahren, in welchem eine pragmatische Fundierung nicht erkennbar ist und aufgrund anders gearteter Anwendungsszenarien offenbar auch nicht benötigt wird. Ziel ist es in *HyTex*, dem selektiven Leser zu einer Verwendungsinstanz eines Terminus nicht *irgendeine* Definition anzubieten, sondern jeweils genau diejenige, die der Termverwendung im aktual rezipierten Textmodul von seiten des Autors (implizit oder explizit) zu Grunde gelegt wurde. Daher wird die pragmatische Typisierung definitorischer Textsegmente dazu genutzt, die Generierung von Linkangeboten zu Termverwendungsinstanzen dahin gehend zu reglementieren, dass im Falle von »Definitionen-Konkurrenz« immer eindeutig entschieden werden kann, welche der konkurrierenden Konzeptualisierungen für eine bestimmte Verwendung eines terminologischen Ausdrucks die höchste Priorität besitzt (und folglich als Zielanker für ein entsprechendes Linkangebot in Frage kommt). Pragmatische Überlegungen

sprechen für die Annahme der folgenden Gewichtungsregeln[2]:

(1) Kommissive Typen >> Assertive Typen
(2) Setzungen >> Fremdzuschreibungen

Regel (1) gründet auf der Annahme, dass kommissiven Typen einen Autor infolge ihres explizit handlungsdeterminierenden Charakters stärker in die Pflicht nehmen als assertive Typen, die primär auf einen Transfer von Wissen gerichtet sind (und überdies – im Gegensatz zu den Kommissiva – falsifizierbar sind). Gleichwohl kann jedoch einer Setzung aufgrund des mit ihr verbundenen universalen Gültigkeitsanspruchs eine ähnlich bindende (wenn auch nicht explizit versprachlichte) Funktion zukommen, sofern sie nicht in Konkurrenz zu einer Selbstzuschreibung oder einer Direktive steht. Konkurriert eine Setzung mit einem kommissiven Typ, so ist der kommissive Typ vorzuziehen; in solchen Fällen gehen wir davon aus, dass (a) wenn die Setzung dem kommissiven Typ vorangeht, mit der Setzung zunächst eine »allgemeine Definition« gegeben wird, während anhand des kommissiven Typs eine Definition gegeben wird, die mit der Setzung zwar kompatibel ist, aber einen höheren Spezifiziertheitsgrad aufweist, und (b) wenn der kommissive Typ der Setzung vorangeht, es sich bei der Setzung lediglich um eine (z.B. didaktisch motivierte) Wiederaufnahme der zuvor qua Selbstzuschreibung oder Direktive eingeführten Definition handelt. Regel (2) ist ergibt sich aus den Gültigkeitsbeschränkungen, die für Fremdzuschreibungen konstitutiv sind.

In Fällen, in welchen zwei Setzungen oder zwei kommissive Typen miteinander konkurrieren, gehen wir davon aus, dass diese hinsichtlich der in ihren Definientes charakterisierten Konzepten kompatibel sind. In Fällen, in welchen zwei Fremdzuschreibungen miteinander konkurrieren (ohne dass zum selben Terminus eine Definition höherrangigen Typs vorliegt), nehmen wir an, dass bei einer der beiden die Angabe der Sprachverwendergruppe bzw. Weltsicht, auf welche die Gültigkeit des definitorischen Sachverhalts eingeschränkt ist, so interpretiert werden kann, dass der Autor der betreffenden Sprachverwendergruppe bzw. Weltsicht zugerechnet werden kann (etwa in folgendem Beispiel einer Konkurrenz zweier definitorischer Fremdzuschreibungen in einer linguistischen Arbeit: »In der Chemie bezeichnet Valenz die Eigenschaft von Elementen, sich mit anderen Elementen zu Molekülen zu verbinden. […] In der Linguistik versteht man unter Valenz die Fähigkeit eines Wortes, andere Wörter semantisch-syntaktisch an sich zu binden.«).

Die beschriebenen Hypothesen werden im *HyTex*-Projekt derzeit anhand einer Auswertung eines repräsentativen Korpusausschnitts empirisch überprüft.

[2] *Erläuterungen*: Regel (1) ist Regel (2) übergeordnet; »A >> B« ist zu lesen als »A hat im Falle zweier oder mehrerer konkurrierender Definitionen zu ein- und demselben Terminus X im Vortext einer Verwendungsinstanz von X höhere Priorität als B«.

Literatur

[Bei02] M. Beißwenger (in Vorb.): Eine handlungssemantische Sicht auf Definitionen in Fachtexten.

[GDS] G. Zifonun, L. Hoffmann & B. Strecker: *Grammatik der deutschen Sprache*. 3 Bde. Berlin. New York 1997 (Schriften des Instituts für deutsche Sprache 7.1-7.3).

[KM01] J. Klavans & S. Muresan: Evaluation of DEFINDER: A System to Mine Definitions from Consumer-Oriented Medical Text. In: *Proceedings of The First ACM+IEEE JCDL* 2001.

[LBS02] E. A. Lenz, M. Beißwenger & A. Storrer: Hypertextualisierung mit Topic Maps – ein Ansatz zur Unterstützung des Textverständnisses bei der selektiven Rezeption von Fachtexten. In: R. Tolksdorf & R. Eckstein (Hrsg.): *XML Technologien für das Semantic Web – XSW 2002 Proceedings*. Bonn 2002 (Lecture Notes in Informatics P-14), 151-159.

[LREC02] M. Gonzàles Rodríguez & C. Paz Suarez Araujo (Eds.): *LREC 2002: Third International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA) 2002.

[LS02] E. A. Lenz & A. Storrer: Converting a corpus into a hypertext: An approach using XML topic maps and XSLT. In: [LREC02], Vol. II, 432-436.

[MK02] S. Muresan & J. Klavans: A Method for Automatically Building and Evaluating Dictionary Resources. In: [LREC02], Vol. I, 231-234.

[PM01] S. Pepper & G. Moore (eds.): *XML Topic Maps (XTM) 1.0*. TopicMaps.Org Specification. http://www.topicmaps.org/xtm/1.0/, 2001.

[Sto02] A. Storrer: Coherence in text and hypertext. In: *Document Design* 3(2). 2002, 156-168.