

Generierung von Linkangeboten zur Rekonstruktion terminologiebedingter Wissensvoraussetzungen

Michael Beißwenger, Eva Anna Lenz, Angelika Storrer

Universität Dortmund, Institut für deutsche Sprache und Literatur
Emil-Figge-Str. 50, D-44227 Dortmund
{beisswenger, lenz, storrer}@hytex.info

Abstract. Dieser Beitrag skizziert Strategien zur (semi-)automatischen Annotation von definitiven 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.

Schlüsselwörter:

Hypertext, Linking, Terminologie, Fachkommunikation, Texttechnologie

1 Projektrahmen und Szenario

Im Projekt "Hypertextualisierung auf textgrammatischer Grundlage" (*HyTex*, vgl. <http://www.hytex.info>) geht es um die (semi-)automatische Generierung von Hypertextsichten, die Nutzern beim Browsen in einer Hypertextumgebung genau diejenigen Wissensvoraussetzungen anbieten, die zum Verständnis des aktuell rezipierten Inhalts benötigt werden. In diesem Kontext werden automatische Strategien zur Segmentierung und zum Linking an einem Fachtextkorpus entwickelt und getestet. Das Korpus enthält in seinem Kernbestand wissenschaftliche Abhandlungen zur Domäne Texttechnologie. Die Strategien zur Generierung von Hypertextsichten operieren über Repräsentationen von Wissen auf dreierlei Ebenen, die in der Veranschaulichung der Projektarchitektur in Abbildung 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 STTS-basierte grammatische Vorannotation aufsetzt.
- Das Wissen über die Relationen zwischen zentralen Konzepten und Termini der Fachtextdomäne wird mit XML Topic Maps (XTM, [PM01]) modelliert (vgl. [LBS02]).
- Annahmen über das Vorwissen bestimmter Nutzertypen werden zunächst als (statische) Nutzerprofile modelliert, in einer späteren 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.

In diesem Beitrag konzentrieren wir uns schwerpunktmäßig 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. 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).

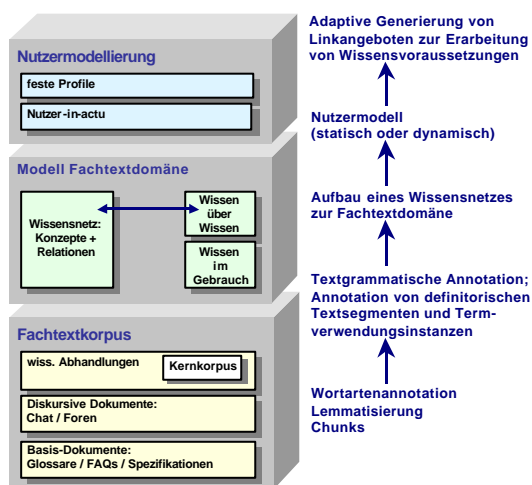


Abb. 1: Die HyTex-Projektarchitektur.

Für die Implementierung unserer Architektur nutzen wir XML als Austauschformat für alle Komponenten: für das Markup des Textkorpora, für die Topic Map (XTM-Syntax) und für die Benutzermodellierung. Dies erlaubt es uns, Dokumente im Web-Kontext zu erstellen und wiederzuverwenden und vorhandene Werkzeuge zu nutzen, z.B. zur Validierung. Zur Erzeugung des späteren Hypertextes aus der textgrammati-

schen Annotation und der Topic Map benutzen wir XSLT; zu Einzelheiten siehe [LS02].

Unter Rückgriff auf Wissen aus den drei Ebenen werden Strategien implementiert, um (a) Hypertext-Sichten auf die Korpus-texte 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; das dazu notwendige Tagset wird in Anlehnung an etablierte Ansätze zur Kategorisierung von sprachlichen Mitteln der Textverknüpfung (Konnektoren, Anaphern) entwickelt.

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 definitivisch eingeführt werden.

2 Strategien zur Annotation von definitivischen Textsegmenten und Termverwendungsinstanzen

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 definitivische Textsegment zu finden ist, welches als Zielanker eines entsprechenden Linkangebots in Frage kommt, müssen nicht nur die Termverwendungsinstanzen, sondern auch sämtliche definitivischen 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 definitivischen Textsegmente im Korpus erfolgt teilautomatisch und auf der Grundlage sowohl einer funktionalen Typologie von Definitionen als auch einer Beschreibung grammatischer Strukturmuster für definitivische Textsegmente mit den im Deutschen als Definitoren verwendeten Prä-

dikatoren (wie z.B. *bezeichnen (als)*, *verstehen (unter)*). 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 *Stoiber bezeichnete die Abstimmung über das Zuwanderungsgesetz als Skandal*), als auch, um anhand der grammatischen und syntaktischen Eigenschaften eines definitorisch und in einer bestimmten flexivischen Ausprägung verwendeten Prädikators automatisch identifizieren zu lassen, in welcher syntaktischen Position des betreffenden Textsegments diejenigen Phrasen zu lokalisieren sind, die im Rahmen der Definition als Definiendum bzw. als Definiens fungieren.

Die für die teilautomatische Annotation von definitorischen Textsegmenten zu Grunde gelegte Typologie von Definitionen unterscheidet auf erster Typologiestufe nach dem Kriterium der Verwendung bzw. des referentiellen Status des als Definiendum fungierenden Terminus (erwähnte vs. nicht-erwähnte bzw. sprachreflexive vs. referentielle Verwendung) in *sprach-* vs. *sachbezogene Definitionen*.

Sprachbezogene Definitionen sind solche Definitionen, die von ihrem Produzenten explizit zu Zwecken der Vermittlung eines Wissens über die Verwendung sprachlicher Ausdrücke gegeben werden und bei welchen die aus dem Definiens erschließbare Information auf ein im Definiendum erwähntes Sprachzeichen bezogen ist. Sprachbezogene Definitionen lassen sich auf einer zweiten Typologiestufe unterteilen in *direkte* und *indirekte sprachbezogene Definitionen*. Der grundlegende Unterschied zwischen *direkten sprachbezogenen Definitionen* einerseits und *indirekten sprachbezogenen* sowie *sachbezogenen Definitionen* andererseits besteht darin, dass letztere falsifizierbar sind, erstere dagegen aufgrund ihres performativen Charakters bestenfalls abgelehnt, nicht aber verneint werden können. Mit einer direkten sprachbezogenen Definition führt ein Fachtextautor ein Ausdruckselement in sein Textuniversum ein und schreibt diesem qua deklarativer Setzung bestimmte Bezugsregeln zu. Er leistet somit (zumindest im Rahmen des jeweils betreffenden Textes) terminologische Systemarbeit gestalterischer Art. Mit einer indirekten sprachbezogenen Definition dagegen wird lediglich eine außerhalb des betreffenden Textes (z.B. von einem anderen Autor oder in einem bestimmten Fachbereich) etablierte Definition referiert und/oder übernommen, ohne dass der Autor selbst als deren Urheber in Erscheinung tritt. Der Unterschied zwischen direkten und indirekten sprachbezogenen Definitionen spielt in unserer Arbeit v.a. für die Ermittlung von Prioritäten bei textinterner "Definitionen-Konkurrenz" (s.o.) eine wichtige Rolle. An folgenden Beispielen lässt sich ersehen, dass den *direkten* gegenüber den *indirekten sprachbezogenen Definitionen* ein höherer Stellenwert beizumessen ist:

- *Werkzeuge, die der Manipulation von Sichten auf der Nutzeroberfläche dienen, nenne ich im Weiteren "Filterwerkzeuge".* [direkte sprachbezogene Definition]
- *Werkzeuge, die der Manipulation von Sichten auf der Nutzeroberfläche dienen, bezeichnet man als Filterwerkzeuge.* [indirekte sprachbezogene Definition]
- *Mey (1996) bezeichnet Werkzeuge, die der Manipulation von Sichten auf der Nutzeroberfläche dienen, als Filterwerkzeuge.* [indirekte sprachbezogene Definition]

Ebenfalls zu den Definitionen rechnen wir bestimmte sachbezogene Aussagen. Sachbezogene Aussagen (also solche Aussagen, in welchen ein Terminus referentiell verwendet wird) sind nicht *per se* definatorisch, können aber in bestimmten Verwendungskontexten *als* Definitionen *fungieren*. Kriterien für das Vorliegen einer *sachbezogenen Definition* sind:

- [*semantische Kriterien*.:] Mit der betreffenden Aussage wird generisch referiert; weiterhin wird mit dem verwendeten Prädikator ein Äquivalenzverhältnis behauptet (z.B. der Form *X ist Y*).
- [*strukturelles Kriterium*.:] Die betreffende Aussage beinhaltet als prädikative Ergänzung eine Phrase, welche die Struktur eines Definiens aufweist (Angabe einer Kategorie in Form einer Nominalphrase plus Angabe eines spezifizierenden Merkmals, in der Regel realisiert durch pränominale AP und/oder postnominale PP und/oder Relativsatz).
- [*kontextuelles Kriterium*.:] An keiner Stelle des Vortextes findet sich eine direkte sprachbezogene Definition mit dem selben Terminus in Definiendumposition.

Beispiele für definatorische Textsegmente aus Fachtexten, denen unter semantischem und strukturellem Aspekt der Status sachbezogener Definitionen zugesprochen werden kann, sind:

- *Links sind computerverwaltete Zuordnungen zwischen Ankern.*
(generische Referenz; Prädikativkonstruktion mit Kopula *sein*; definiensartige Prädikativergänzung: [NP [AP computerverwaltete] Zuordnungen [PP zwischen Ankern]])
- *Ein Meson ist ein zusammengesetztes Teilchen, das aus je einem Quark und einem Antiquark besteht.*
(generische Referenz; Prädikativkonstruktion mit Kopula *sein*; definiensartige Prädikativergänzung: [NP ein [AP zusammengesetztes] Teilchen [S das ... besteht]])

Zu den beschriebenen Typen von Definitionen lassen sich jeweils weitere Subtypen angeben.

Bei "Definitionen-Konkurrenz" gehen wir davon aus, dass direkten sprachbezogenen Definitionen (aufgrund ihres performativen Potentials) ein höherer Stellenwert zuzusprechen ist als indirekten sprachbezogenen und sachbezogenen Definitionen. Regeln für die Gewichtung von indirekten sprachbezogenen gegenüber sachbezogenen Definitionen können hingegen erst auf der Grundlage empirischer Untersuchungen am textgrammatisch annotierten Korpus sicher festgelegt werden.

Literatur

- [LBS02] Eva Anna Lenz, Michael Beißwenger & Angelika Storrer (erscheint): Hypertextualisierung mit Topic Maps. In: *XML Technologien für das Semantic Web (XSW 2002)*.
- [LS02] Eva Anna Lenz & Angelika Storrer (erscheint): Converting a corpus into a hypertext: An approach using XML topic maps and XSLT. In *LREC 2002: Third international conference on language resources and evaluation*, 2002.
- [PM01] Steve Pepper and Graham Moore (eds.): XML Topic Maps (XTM) 1.0. TopicMaps.Org Specification. <http://www.topicmaps.org/xtm/1.0/>, 2001.