

Annotation definatorischer Textsegmente und »terminologiesensitives Linking«

Arbeitsbericht

Projekt
**Hypertextualisierung auf
textgrammatischer Grundlage**
(www.hytex.info)

Michael Beißwenger

2004

- 1 Terminologisches Wissen und seine Explizierung als Schlüssel für das Verständnis von Fachtexten und fachsprachlicher Kommunikation
- 2 Annotation/Typisierung von Definitionen und Termverwendungsinstanzen und „terminologiesensitives Linking“ – der pragmatische Ansatz des *HyTex*-Projekts
 - 2.1 *Schritt 1*: Manuelle Annotation von definatorischen Textsegmenten
 - 2.2 *Schritt 2*: Erzeugung einer Termliste und automatische Annotation von Termverwendungsinstanzen
 - 2.3 *Schritt 3*: Typisierung von definatorischen Textsegmenten auf Basis einer pragmatischen Typologie definatorischen Sprachhandelns
 - 2.4 *Schritt 4*: Erarbeitung von Ranking-Regeln für das automatische Linking
- 3 Literatur

1. Terminologisches Wissen und seine Explizierung als Schlüssel für das Verständnis von Fachtexten und fachsprachlicher Kommunikation

Ein wesentliches (genetisches wie lexikalisches) Merkmal von Fachsprachen ist die Einführung und Verwendung von Termini. Im Gegensatz zu lexikalischen Einheiten der Alltagssprache ist die Verwendung von Termini nicht qua stillschweigender Übereinkunft „eingespielt“ (und somit in Hinblick auf die Grenzen ihrer Verwendung latent variierbar); vielmehr sind Termini in der Regel per Definition auf ein ganz bestimmtes und klar fixiertes Konzept festgelegt:

Im Unterschied zum Wandel von alltäglichen Benennungen geschieht die Veränderung von Termini nicht (sensu Keller 1995) von unsichtbarer Hand. (Wiegand 1996: 95)

Termini gewährleisten somit im Gegensatz zu gemeinsprachlichen Ausdrücken die Möglichkeit einer präzisen Bezugnahme auf Konzepte und fachliche Gegenstände. Weiterhin erlauben sie im Falle komplexer und abstrakter Konzepte eine Ökonomisierung fachlicher Kommunikationsprozesse: Wenn ein Konzept einmal ausführlich charakterisiert und per Festsetzung (Definition) mit einem Terminus „etikettiert“ wurde, so können Experten der betreffenden Fachdomäne dieses Etikett fortan verwenden, um im Rahmen ihres Austauschs eindeutig auf das betreffende Konzept bezug zu nehmen, ohne dieses Konzept jeweils erneut explizieren zu müssen (vgl. z.B. Gorski 1967: 432). Ziel „terminologischer Praxis“ (also der Verständigung anhand von Termini) ist es,

Prozesse der fachlichen Verständigung (Fachkommunikation) in ihrer Effizienz, Eindeutigkeit, Präzision und Klarheit zu unterstützen (Budin 2000: 5).

Voraussetzung ist natürlich, dass die am Kommunikationsprozess Beteiligten sämtlich über das entsprechende Vorwissen (also das Wissen um das einem Terminus qua Definition zugeordnete Konzept) verfügen.

Eine idealisierte Sichtweise auf fachliches „Expertentum“ müsste folglich davon ausgehen, dass innerhalb einer Fachdomäne diejenigen Wissenschaftler zu den „Experten“ zu zählen sind, die mit dem relevanten domänenspezifischen Fachwortschatz souverän vertraut sind (d.h.: die relevanten Konzepte sowie die zugehörigen Termini kennen und in Produktions- wie Rezeptionssituationen kompetent anwenden können). Zur Gruppe der „Semi-Experten“ wären hingegen all diejenigen produktiv und/oder rezeptiv am Diskurs innerhalb der Domäne teilhabenden Personen zu rechnen, auf welche dies nur teilweise zutrifft.

Misst man in Anbetracht der kontinuierlichen Ausdifferenzierung von Fachdomänen und Wissenschaftsdisziplinen und dem damit einhergehenden rasanten Zuwachs spezialisierter Fachpublikationen eine solche idealisierte Sichtweise an der „Wissenschaftswirklichkeit“, so ist festzustellen, dass selbst innerhalb der Grenzen einer Fachdomäne letztlich nur sehr wenige Forscher zur Gruppe der „Experten“ gezählt werden können. Kaum ein Wissenschaftler hat heute mehr die Ressourcen, sämtliche für seinen Fachbereich relevanten Neuerscheinungen eingehend zu studieren und sämtliche innerhalb der „sozialen Welt“ seiner Domäne laufenden Kommunikationsprozesse (Fachtagungen, Vorträge, Forschungsprojekte) zeitnah zu verfolgen. In Anbetracht der zusätzlich gegebenen Notwendigkeit, für die eigene Arbeit immer auch partiell die Entwicklungen und Diskussionen in angrenzenden Fachdomänen zu verfolgen, ist letztlich der „ideale Experte“ in den verschiedenen Bereichen seiner Tätigkeit und seines Interesses immer nur ein „Semi-Experte“, der sich darum bemühen muss, zumindest die wichtigsten Entwicklungen und Neuerscheinungen auf seinem primären Gebiet wie auch auf den diesem Gebiet benachbarten Gebieten zu verfolgen. Die humanistische Maxime „Man muss

auf einem Gebiet Fachmann sein, auf allen anderen Gebieten zumindest Dilettant“ erweist sich im Informationszeitalter mehr denn je als ein Ideal, das zwar nach wie vor erstrebenswert ist, für eine Einzelperson aber immer nur mit diversen Abstrichen erreicht werden kann.

Wo der Experte sich immer mehr als „Semi-Experte“ erweist, werden die „idealen Semi-Experten“ zugleich immer mehr zu „Fach-Laien“: Personengruppen, die gemäß ihrer Aufgaben- und Tätigkeitsfelder eher rezeptiv und reproduktiv als produktiv und innovativ an Kommunikationsprozessen innerhalb von Fachdomänen teilhaben (z.B. Studierende, Journalisten, Forscher aus Nachbardisziplinen), sind darauf angewiesen, in zumeist kurzer Zeit sich einen Überblick über zentrale Konzepte und Diskussionslinien einer Domäne zu verschaffen, um diesen dann im eigenen Tätigkeitsbereich verwerten zu können. Im Zuge der Ausdifferenzierung von Disziplinen und den kontinuierlich steigenden Zuwachsraten an Publikationen wird aber auch diese Teilhabe an Wissenschaftsbereichen zunehmend schwieriger, zumal aufgrund der oftmals komplexen intertextuellen Bezüge zwischen fachlichen Textbeiträgen sowie aufgrund komplex ausdifferenzierter und z.T. auch schulenabhängiger Fachwortschätze und Terminologiesysteme das adäquate Verständnis eines Textes nicht selten die Kenntnis einer Anzahl anderer Texte erfordert und die Vertrautheit mit einem bestimmten Ausschnitt aus der in der Domäne verwendeten Terminologie voraussetzt.

Fluck (1996) beschreibt das Kommunikationsproblem innerhalb von Fachdomänen und bei der Konfrontation mit fachsprachlichen Textsorten wie folgt:

Die Existenz und das permanente Anwachsen der Fachsprachen ist heute zu einem Kommunikationsproblem ersten Ranges geworden. Immer wieder liest und hört man, daß es zunehmend schwieriger und oft unmöglich wird, zwischen den verschiedenen Bevölkerungsschichten, zwischen Fachleuten und Laien und zwischen den Fachleuten untereinander sich zu verständigen. Diese Sprachnot kennt der Wissenschaftler ebenso wie der Techniker, der Übersetzer oder der Journalist. Sie hat inzwischen ebenfalls zu einer Informationsbarriere geführt, die sich beispielsweise auf dem Gebiet der Wissenschaftspublizistik deutlich abzeichnet. [...] Nicht nur für den Laien, auch für die Wissenschaftler selbst erweisen sich die herausgebildeten, differenzierten Fachsprachen oft als unüberwindbares Hindernis. Ein Biologe wird einen Juristen, ein Chemiker einen Soziologen kaum mehr verstehen können. Und das Aneinandervorbeireden ist heute selbst unter Wissenschaftlern eines einzigen Faches keine Seltenheit mehr. (Fluck 1996: 37)

Darüber hinaus werden Fachtexte zumeist selektiv rezipiert: Wer einen Fachtext „liest“, tut dies häufig unter einer spezifischen Fragestellung bzw. mit einem spezifisch gearteten Informationsbedarf und durchstreift den Text daher eher im Modus des „suchenden Lesens“ (per Zugriff via Inhaltsverzeichnis, Schlagwortregister oder „Herumblättern“) als ihn intensiv und mit dem dafür nötigen Zeitkontingent von Anfang bis Ende durchzuarbeiten und sämtliche Gedankengänge des Autors nachzuvollziehen. Die Folge aus solchen (einer Bewältigung der Publikationsflut geschuldeten) Rezeptionsstrategien sind (neben dem ohnehin gegebenen Problem des schulen-, teildisziplinen- und bisweilen auch autorenabhängigen Sprachgebrauchs) weitere terminologiebedingte Verständnisprobleme, die sich daraus ergeben, dass das Lesermodell des Autors (nach welchem der „ideale Adressat“ den Text von Anfang bis Ende liest) den tatsächlichen Rezeptionsgepflogenheiten eines Großteils seiner Adressatengruppe (die den Text „querlesen“ und in ihm nach genau derjenigen Information suchen, die sie in ihm zu entdecken hoffen) zuwiderläuft. Ein „querlesender“ Rezipient kann, wenn er einer Verwendungsinstanz eines terminologischen Ausdrucks begegnet, nicht mit Sicherheit annehmen, dass dieser Ausdruck eine terminologische Einheit repräsentiert, die disziplinen-, schulen-, autor- oder textspezifisch ist (es sei denn, er ist beim „Querlesen“ bereits einer

Textpassage begegnet, in welcher eine Konzeptualisierung zum betreffenden Ausdruck explizit eingeführt wurde). Ein „querlesender“ Rezipient hat somit in Bezug auf einen für ihn nicht oder nicht exakt semantisierbaren sprachlichen Ausdruck zu entscheiden,

1. ob es sich bei dem betreffenden Ausdruck um einen Terminus handelt oder nicht¹;
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.

Im Grunde wäre es daher ratsam, wenn Fachautoren an jeder Stelle ihrer Texte, an welcher sie einen Terminus verwenden, eine Fußnote anfügen, in welcher die ihrem spezifischen Gebrauch dieses Terminus zugrunde gelegte Definition oder zumindest ein Verweis auf eine Referenzstelle angegeben würde. In Anbetracht der Kosten von Druckerzeugnissen sowie aus Gründen der Übersichtlichkeit ist dies natürlich eine illusorische Vorstellung.

Prinzipiell kann ein Fachautor in einer Passage seines Fachtextes einen Terminus entweder *verwenden* oder *definieren*:

Im ersten Fall setzt der Verfasser voraus, daß der Terminus alleine ausreicht, um beim Adressaten die Wissensmenge zu evozieren, die der Terminus im Text vertritt. Der Terminus dient im Text in diesem Falle als Platzhalter für eine Wissensmenge, die der Leser selbst aus seinem Vorwissen abrufen soll. Im zweiten Fall wird umgekehrt angenommen, der Terminus könne beim Adressaten diese Wissensmenge eben nicht evozieren. (Kastberg 1999: 44)

Welche der beiden Optionen ein Autor wählt, hängt von verschiedenen Vorentscheidungen ab:

- (a) Der Autor entscheidet sich an einer Textstelle t_j für die *Verwendung* eines Terminus T_i , weil
 - (a-1) er T_i an einer im linearen Verlauf seines Textes vorausgehenden Textstelle t_i bereits definiert hat (und weil er – wie die meisten Fachautoren – bei der Textproduktion von einem LesermodeLL ausgeht, das einen „idealen Fachtextleser“ anvisiert, der seinen Text von Anfang bis Ende durchliest),
 - (a-2) er den Terminus in einer Art und Weise verwendet, die den Verwendungsgewohnheiten in der Fachdomäne oder zumindest in der Schule, welcher der Autor zugehört, entspricht und er als Adressaten seines Textes Personen anvisiert,

1 Beispielsweise wird ein Rezipient aus einer Nachbarwissenschaft nicht unbedingt auf die Idee kommen, dass, wenn er eine linguistische Arbeit querliest, es sich bei einem Vorkommen des Ausdrucks „Satz“ um die Verwendung eines terminologischen Ausdrucks handelt, dessen Konzeptualisierung in der betreffenden Disziplin hochgradig schulen- und/oder autorabhängig ist. Da der Ausdruck „Satz“ auch in der Gemeinsprache lexikalisiert ist (dort aber in weitaus weniger Anlass zur Diskussion gebender Art und Weise als in der Linguistik), könnte unser nicht-linguistischer Leser vielmehr vermuten, dass es sich beim Vorkommen des Ausdrucks „Satz“ im rezipierten Fachtextausschnitt gar nicht um einen terminologischen Ausdruck handelt und daher seine Alltagskonzeptualisierung zu „Satz“ auf die Semantisierung der betreffenden Textpassage anwenden (was dann unter Umständen zu einer teilweisen oder vollständigen Missinterpretation der betreffenden Textpassage führen kann).

die in der Domäne oder zumindest in der betreffenden Schule Expertenstatus haben.

- (b) Der Autor entscheidet sich an einer Textstelle ξ für die *Definition* eines Terminus T_i , weil
- (b-1) er T_i an Textstelle ξ zum ersten Mal verwendet und er nicht voraussetzen kann oder will, dass die Adressaten seines Textes mit dem in der Domäne oder Schule üblichen Gebrauch von T_i vertraut sind,
 - (b-2) er T_i im folgenden Text auf der Grundlage einer vom in der Domäne oder Schule üblichen Gebrauch abweichenden Konzeptualisierung verwenden möchte.

Ob ein Autor innerhalb seines Textes eine Definition zu einem von ihm verwendeten Terminus gibt oder nicht, kann auch mit einer fach- oder schuleninternen „Geschichtlichkeit“ von Termini in Zusammenhang stehen:

Wenn beispielsweise ein Sprachwissenschaftler eine spezifische Fragestellung bearbeiten möchte, die zur Referenzsemantik gerechnet werden kann, dann ist er angesichts der zahlreichen, miteinander konkurrierenden Auffassungen davon, was unter Referenz zu verstehen ist, genötigt anzugeben, was er unter Referenz versteht, wenn er sich hinreichend verständlich machen will. Denn die Vorverständigung über den Begriff der Referenz im Fach insgesamt ist nur von sehr genereller Art und reicht über eine Charakterisierung von der Art wie: „Wenn von Referenz die Rede ist, geht es um den Bezug einer Sprache zur Welt“ nicht (oder nicht weit) hinaus; alle weiteren Präzisierungen sind theoriespezifisch. Dies bedeutet nun, daß der Wissenschaftler einerseits hinsichtlich des Gebrauchs von *Referenz* an die wissenschaftliche Sprachtradition des Faches gebunden ist, andererseits jedoch erhebliche Freiheitsgrade für seinen Gebrauch von *Referenz* hat. (Wiegand 1996: 95)

Auf jeden Fall hat jeder Fachautor (im Gegensatz zu gemeinsprachlichen Sprachverwendern) die Freiheit, seine Sprache selbst zu „machen“:

Wissenschaftler sind in einem relativ weitgespannten Rahmen Herr des terminologischen Lexikons. Sprachliche Setzungen sind gang und gäbe, systembezogene Nominations- und Umbenennungshandlungen treten häufig auf, der lexikalische Wandel insbesondere im Bereich der Terminologien vollzieht sich aufgrund der gestalterischen Eingriffe sichtbarer Wissenschaftlerhände, und das Einordnen der eigenen Texte in die thematisch einschlägigen Texte der scientific community ist ein Procedere, das notwendiger Teil des gesellschaftlichen Verfahrens der Erkenntnisgewinnung ist. (Wiegand 1996: 98)

Letztlich kann ein- und derselbe Fachautor sogar in unterschiedlichen Schaffensabschnitten seiner wissenschaftlichen Biographie ein- und denselben Terminus unterschiedlich verwenden oder aber ein- und dasselbe Konzept mit unterschiedlichen terminologischen Etikettierungen versehen.

Beispiel:

In früheren Arbeiten habe ich für die Gesamtstruktur eines Wörterbuches den Terminus *Hyperstruktur* verwendet. [...] Der Vorschlag wurde – soweit ich weiß – weder kritisiert noch aufgegriffen. Inzwischen ist er obsolet. Der Terminus ist jetzt auch rein sprachlich ungeeignet, und zwar wegen solcher inzwischen akzeptierter Termini wie *Hypertext*. [...] Es ist inzwischen wohl akzeptiert, daß ein Wörterbuch, in texttheoretischer Perspektive betrachtet, als ein Textverbund gelten kann. [...] Da *Textverbund* ein Terminus ist, der in texttheoretischer Perspektive ein Wörterbuch in seiner textuellen Gesamtheit meint, gibt es auch die Möglichkeit, *Textverbundstruktur*

als Terminus zu verwenden. [...] Ich schlage daher vor, die Gesamtstruktur eines Wörterbuches mit dem Terminus *Textverbundgesamtstruktur* zu bezeichnen.²

Das Beispiel zeigt, dass ein Autor in ein- und demselben Fachtext einander widersprechende Definitionen zu ein- und demselben Terminus versprachlichen kann, ohne dabei Gefahr zu laufen, in Hinblick auf die von ihm verwendete Fachsprache inkonsistent zu werden. Während der Autor des Beispiels in zweiterer Definition deutlich macht, dass die darin explizierte Zuordnung zwischen einem Konzept und dem terminologischen Ausdruck *Textverbundgesamtstruktur* seiner aktuellen Sprachverwendung entspricht, verweist er mit ersterer Definition, in welcher dasselbe Konzept einem anderen terminologischen Etikett zugeordnet wird, darauf, dass diese Definition zwar von ihm selbst stamme, allerdings eine Sicht auf die sprachliche Fassung des betreffenden Fachgegenstandsausschnitts repräsentiere, die in der Vergangenheit angesiedelt ist. An Definition Nummer 1 selbst ist jedoch noch nicht abzulesen, dass diese vom Autor zum Zeitpunkt der Abfassung seines Textes nicht mehr vertreten wird. In Zusammenschau mit Definition Nummer 2, die eine widersprüchliche Terminologisierung des betreffenden Konzepts vornimmt, zeigt sich jedoch, dass Definition 1 wohl kaum diejenige sein kann, die für die Benennung des betreffenden Konzepts im weiteren Text relevant ist, sondern dass Definition 2 als diejenige zu erachten ist, die maßgeblich ist für die sprachliche Fassung des betreffenden Konzepts im weiteren Verlauf des vorliegenden Textes. Auf die unterschiedlichen Typen von Definitionen (einerseits Zuschreibung definitorischer Zuordnungen zu anderen „Weltsichten“, andererseits explizite sprachliche Markierung der Verbindlichkeit einer definitorischen Zuordnung für die Sprachverwendung im weiteren Textverlauf) wird an späterer Stelle noch genauer einzugehen sein.

Zwischenfazit: Das Problem der Orientierung in Fachdomänen und des adäquaten Verständnisses von Fachtexten und fachsprachlicher Kommunikation gründet darin, dass

- (a) Fachtexte vor allem selektiv rezipiert werden;
- (b) Fachtexte von ihren Autoren jedoch in aller Regel strikt sequenziell strukturiert werden (konzipiert für einen Rezipienten, der sie von Anfang bis Ende liest und daher an jeder Stelle des Textes dasjenige Wissen im Zugriff hat, welches im Vortext bereits expliziert wurde);
- (c) Fachterminologie, so wie sie in Texten verwendet wird, als hochgradig *autorspezifisch*, oftmals (bei einem Vergleich unterschiedlicher Texte desselben Autors) sogar als *textspezifisch* zu gelten hat.

2 Das Beispiel ist (mit geringfügigen Modifikationen) einer metalexikographischen Arbeit von Herbert Ernst Wiegand entnommen.

2. Annotation/Typisierung von Definitionen und Termverwendungsinstanzen und „terminologiesensitives Linking“ – der pragmatische Ansatz des Hy- Tex-Projekts

Um Leser mit „Semi-Experten“-Status bei der Bewältigung terminologiebedingter Verständnisprobleme bei der selektiven Rezeption digital verfügbarer Fachtexte bestmöglich zu unterstützen, sind für die Designer entsprechender Rezeptionsumgebungen zwei Teilaufgaben zu bewältigen:

Teilaufgabe 1:

eine weitestmöglich automatische Identifizierung und Extraktion bzw. Annotation definitorischer Textsegmente in den betreffenden Texten;

Teilaufgabe 2:

im Falle mehrerer zu einem Terminus gefundener Definitionen die Ermittlung, welche dieser Definitionen für die Verwendung des Terminus an einer ganz bestimmten Stelle in einem ganz bestimmten Text relevant ist.

Teilaufgabe 1 zielt darauf, definatorische Textsegmente überhaupt zu identifizieren, z.B. als Ressource für den Aufbau von Glossaren oder Termdatenbanken (und natürlich auch als Grundlage für die darauf aufbauende Teilaufgabe 2). Teilaufgabe 2 zielt darauf, einem selektiv zugreifenden Rezipienten an derjenigen Textstelle, welche er jeweils gerade rezipiert, genau diejenige Definition als Verständnishilfe anbieten zu können, die für die adäquate Konzeptualisierung einer konkreten Verwendungsinstanz eines Terminus relevant ist. Teilaufgabe 2 zielt also auf die Unterstützung bei der Lösung eines konkreten, textstellenbezogenen Rezeptionsproblems – einen „Semi-Experten“ würde es eher verwirren, böte man ihm bei der Lektüre eines Textmoduls (z.B. über ein Glossar) jeweils sämtliche, sich z.T. widersprechende, Definitionen an, die sich zu einem Terminus in der Dokumentenbasis auffinden lassen. Teilaufgabe 1 ist insbesondere eine computerlinguistische Herausforderung: Auf der Grundlage geeigneter POS-Tagger und Chunk-Parser müssen Suchstrategien entwickelt, implementiert und evaluiert werden, die es erlauben, mit hohen Precision- und Recall-Ergebnissen solche Textsegmente aufzufinden, in welchen Konzeptualisierungen zu Termini versprachlicht werden. Diese Aufgabe ausschließlich auf der Grundlage syntaktischer Indikatoren bewältigen zu wollen, ist insofern problematisch, als die Vielfalt an Versprachlichungsmöglichkeiten für Definitionen enorm groß ist:

Unfortunately there are so many ways in which definitions are conveyed in natural language that it is difficult to come up with a full set of linguistic patterns to solve the problem. (Saggion 2004: 1)

Neuere Ansätze, die sich einer automatischen Auffindung von definatorischen Textsegmenten widmen, beziehen daher neben syntaktischen Suchmustern und ggf. einem Inventar an Verben, die potenziell als definatorische Prädikatoren fungieren können, weitere – text-externe – Daten mit ein, z.B. vordefinierte Terminlisten zu Gegenstandsbereichen oder Fachdomänen sowie Informationen zur lexikalischen Strukturiertheit von Wortschätzen (letztere beispielsweise, um zu ermitteln, ob ein Terminus innerhalb eines Satzes zusammen mit einem Hyponym auftritt; vgl. Saggion 2004).

Eine weitere Herausforderung von Teilaufgabe 1 besteht darin, dass eine Identifizierung von definatorischen Textsegmenten nicht ausschließlich auf der Satzebene operieren kann, sondern

auch die Satzgrenze überschreitende Vertextungsprozeduren (insbesondere Koreferenzbezüge) mit berücksichtigen muss. Ein Satz wie der folgende erfüllt zwar alle Kriterien einer Definition, ist aber für einen Nutzer nicht hilfreich, wenn es darum geht, sich terminologisches Wissen in bezug auf den Ausdruck *1:1-Link* anzueignen:

Beispiel:

Verknüpfungen, die die vorgenannten Bedingungen erfüllen, bezeichne ich als 1:1-Links.

In der ersten Phase des *HyTex*-Projekts war zunächst angedacht, Strategien zum Auffinden und zur automatischen Annotation von definitorischen Textsegmenten in Fachtexten zu entwickeln. Dies erschien insbesondere deshalb als attraktiv, da vergleichbare Strategien bislang lediglich für das Englische existieren – etwa der DEFINDER-Ansatz (vgl. z.B. Klavans & Muresan 2001, Muresan & Klavans 2002) oder der in Saggion (2004) beschriebene Ansatz –, nicht aber für das Deutsche. Um die Herausforderungen auszuloten, die sich für die Entwicklung entsprechender Strategien stellen, wurden zunächst von Mitarbeitern des Projekts „Deutsches Referenzkorpus“ (DEREKO, <http://www.sfs.nphil.uni-tuebingen.de/dereko/>) am Seminar für Sprachwissenschaft der Universität Tübingen 63 Dokumente aus dem *HyTex*-Projektkorpus anhand des Werkzeugs *KaRoPars* (v.0.36)³ um eine Annotation von Wortarten (POS), Chunks und topologischen Feldern angereichert. Die Grundlage für das Wortarten-Tagging bildete dabei das „Stuttgart-Tübingen-Tagset“ (STTS, <http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html>). Der hierbei gewonnene linguistisch aufbereitete Korpusausschnitt sollte – in Kombination mit einer manuell erstellten Termkandidatenliste für die Fachdomäne sowie einer in der Aristoteles-Tradition stehenden Annahme über den strukturellen Aufbau von Definitionen – die Datenbasis für die Implementierung und Evaluation syntaktischer Suchmuster bilden. Parallel dazu wurde in Kooperation mit der Firma *TEMIS* (<http://www.temis-group.com/>) mit dem Text-Mining-Werkzeug *Knowledge Extractor* anhand eines kleinen Testkorpus ausgetestet, wie zielführend ein Vorgehen zur Identifizierung definitorischer Textsegmente sein kann, das eine Termkandidatenliste in Kombination mit syntaktischen Suchmustern nutzt. Perspektivisch war angedacht, den *Knowledge Extractor* (der standardmäßig einen integrierten POS-Tagger verwendet und zu analysierende Textdaten „on the fly“ annotiert) auf das STTS-/DEREKO-Tagset anzupassen, um ihn dann als Schnittstelle für die Entwicklung und Verfeinerung von Suchmustern über dem STTS-/DEREKO-annotierten Korpusausschnitt nutzen zu können.

Die Ergebnisse dieser Pilotstudie zeigten letztlich, dass die Vielfalt der Vertextungsmöglichkeiten von Definitionen derart breitgefächert ist, dass eine befriedigende Bearbeitung von Teilaufgabe 1 neben einer Termliste, einer Grundstruktur von Definitionen sowie darauf abgestimmten syntaktischen Suchmustern weitere Ressourcen mit einbeziehen müsste, nämlich insbesondere (a) Ressourcen über die lexikalische Geordnetheit der terminologischen Einheiten der Fachdomäne (ein terminologisches Wortnetz) und (b) Annotationen zu Koreferenzbezügen innerhalb der Korpusdokumente. Da der Aufbau solcher Ressourcen selbst Teil des *HyTex*-Projekts war, hätte die Entwicklung einer Prozedur zur automatischen Identifizierung von definitorischen Textsegmenten (inklusive der Programmierung von Schnittstellen, die es erlauben, die Daten aus *TermNet* und aus der Ebene der Annotation von Koreferenzbezügen miteinzubeziehen) erst am Ende des Projekts stehen können. Prinzipiell ist angedacht, die

3 Die Annotationsstruktur des von *KaRoPars* generierten Outputs ist im zugehörigen „Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z)“ von Frank Henrik Müller (2004) beschrieben (siehe <http://www.sfs.uni-tuebingen.de/~fhm/Biblio/stylebook-04.pdf>).

Bearbeitung dieser Aufgabe zum Gegenstand eines Folgeprojekts zu machen, welches die im Rahmen von *HyTex* erarbeiteten Annotationen auf mehreren Ebenen sowie das in XTM repräsentierte *TermNet* als Ausgangsbasis nutzt.

In Anbetracht der Tatsache, dass ein zentraler Fokus der Projektarbeit auf der Erarbeitung von Strategien zur Unterstützung der selektiven Fachtextrezeption und damit zusammenhängend von Strategien für das automatische Linking lag, wurde für den Aufgabenbereich „Terminologiebedingte Verständnisprobleme“ das Hauptaugenmerk auf Teilaufgabe 2 gelegt, also auf die Entwicklung eines Ansatzes zur automatischen Ermittlung, welche Definition für eine je konkrete Termverwendungsinstanz die relevante ist. Hierfür wurde davon ausgegangen, dass sich sämtliche im Korpus enthaltenen definitorischen Textsegmente bereits im Zugriff befinden. Entsprechend wurden in einem ersten Schritt in den vier „Pilottexten“ sämtliche definitorischen Textsegmente von Hand annotiert (siehe nachfolgend Abschnitt 2.1). Anschließend wurden sämtliche Termini, die in den gefundenen definitorischen Textsegmenten in Definendum-Position auftraten, in eine Termliste überführt. Nach systematischer Ergänzung dieser Liste um die flexionsmorphologischen Varianten der enthaltenen Termini wurden sämtliche Verwendungsinstanzen der Termini in den vier Texten dann automatisch annotiert und auf ihre jeweilige Grundform zurückgeführt (vgl. 2.2). Die Termliste bildete – zusammen mit den Definiens der in Schritt 1 annotierten Definitionen – darüber hinaus die Basis für den Aufbau eines *terminologischen Wortnetzes* (*TermNet*, vgl. Beißwenger, Storrer & Runte 2004). Auf Basis einer an Zweckbereichen sprachlichen Handelns orientierten Typologie definitorischen Sprachhandelns wurden den einzelnen in Schritt 1 ausgezeichneten definitorischen Textsegmenten in der Folge Sprachhandlungstypen zugewiesen (vgl. 2.3). Zuletzt wurde für Fälle von „Definitionen-Konkurrenz“ (d.h.: für Fälle, in welchen zu einer Termverwendungsinstanz mehrere Definitionen des Terminus im Vortext existieren) ein Inventar an Regeln formuliert, welches die einzelnen definitorischen Sprachhandlungstypen unter Einbezug der Reihenfolge ihres Auftretens im linearen Textverlauf pragmatisch gegeneinander gewichtet. Diese Regeln bildeten die Grundlage für die Implementierung eines Ranking-Algorithmus, der es erlauben sollte, zu jeder Termverwendungsinstanz automatisch einen Link auf diejenige Definition zu generieren, die in pragmatischer Hinsicht als am verbindlichsten für die betreffende Verwendungsinstanz gewertet werden kann (vgl. 2.4).

2.1 Schritt 1: Manuelle Annotation von definitorischen Textsegmenten

Annotiert wurden lediglich Definitionen im engeren Sinne, d.h.: solche Textsegmente, in welchen entweder explizit einem terminologischen Ausdruck ein Konzept zugeordnet wird (*Nominaldefinitionen*) bzw. Äquivalenz zwischen dem Referenten eines terminologischen Ausdrucks und einer Beschreibung einer Objektklasse behauptet wird (*Realdefinitionen*). Definitionen des ersteren Typs sind *sprachbezogen*, insofern durch Verwendung entsprechender Prädikatoren (z.B. *bezeichnen als*, *verstehen unter*) deutlich gemacht wird, dass mit der definitorischen Zuordnung eine Zeichenverbindung (Ausdruck–Inhalt) etabliert bzw. beschrieben wird. Definitionen des zweiten Typs sind *sachbezogen*, da in ihnen auf dem Wege einer Äquivalenzbehauptung, die sich auf die Objektwelt bezieht, auf zwei unterschiedliche Weisen auf ein- und dieselbe Klasse von Dingen referiert wird: einmal generisch anhand eines terminologischen Ausdrucks, einmal in Form einer Beschreibung von Eigenschaften, welche für die betreffende Klasse klassenbildend sind. Beide Arten der Definition sind jedoch dadurch motiviert, ein Konzeptwissen zu vermitteln, auch wenn sie sich im gewählten Verfahren unterscheiden: *Nominaldefinitionen* vermitteln Konzeptwissen sprachreflexiv, indem sie zu ei-

nem terminologischen Ausdruck eine Charakteristik des ihm zeichenhaft verbundenen Inhalts angeben; *Realdefinitionen* konstatieren (objektsprachlich) die Identität von verschiedenen Gegenständen bzw. Konzepten mit verschiedenen Namen:⁴

Nominaldefinition (sprachbezogen):

Unter einem *Link* verstehe ich eine computerverwaltete Zuordnung zwischen Ankern.
Computerverwaltete Zuordnungen zwischen Ankern werden als *Links* bezeichnet.

Realdefinition (sachbezogen):

Links sind computerverwaltete Zuordnungen zwischen Ankern.

Links, also computerverwaltete Zuordnungen zwischen Ankern, (...)

Bei der Annotation nicht berücksichtigt wurden andere, nicht-definitorische Arten von Terminologisierungsoperationen, die im Bereich terminologischer Praxis ebenfalls eine Rolle spielen, nämlich Verfahren der *Lokalisierung*, der *Umterminologisierung* und der *Terminologieübernahme*.

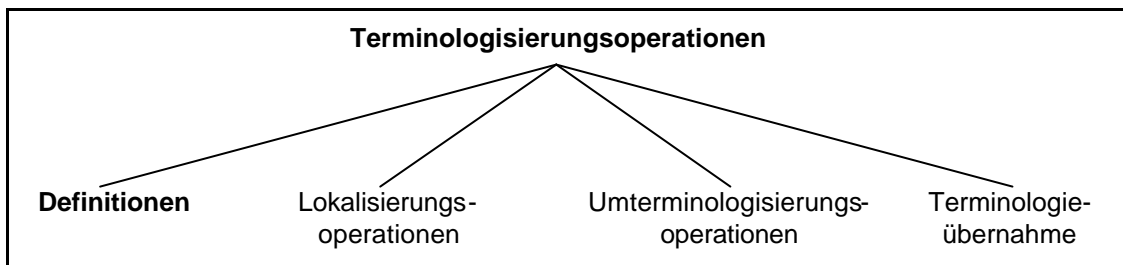


Abb. 1: Typen von Terminologisierungsoperationen.

Bei *Lokalisierungsoperationen* wird einem terminologischen Ausdruck nicht ein Konzept zugeordnet, sondern der Ausdruck wird in Beziehung gesetzt zu Ausdrücken aus einer anderen Sprache, die mit dem gleichen Konzept verbunden sind.

Beispiel:

In der englischsprachigen Literatur werden Linkanzeiger häufig *buttons*, *icons* oder *hotwords* genannt.

Bei *Umterminologisierungsoperationen* wird einem terminologischen Ausdruck nicht ein Konzept zugeordnet, sondern ein (bereits – z.B. bei einem anderen Autor – existenter) terminologischer Ausdruck wird durch einen alternativen terminologischen Ausdruck ersetzt. Es wird also nicht konzeptualisiert, sondern lediglich eine Ausdruckseinheit (unter Beibehaltung des Konzepts) durch eine andere ersetzt.

Beispiele:

Den von (Tochtermann 1995) benutzten Terminus „Verweis“ werde ich durch den Terminus „Link“ ersetzen, weil das Verweiskonzept im gedruckten Medium gerade vom Link-Konzept in Hypertext abgehoben werden soll.

Der von (Tochtermann 1995) benutzte Terminus „Knoten“ soll wegen der besseren Wortbildungsmöglichkeiten (z.B. modular, Modularisierung) durch den Ausdruck „Modul“ ersetzt werden.

Bei der *Terminologieübernahme* wird einem terminologischen Ausdruck nicht ein Konzept zugeordnet, sondern es wird lediglich ein Ort benannt, an welchem diejenige Konzeptualisie-

4 Zur Unterscheidung von Nominal- und Realdefinitionen siehe z.B. Gorski (1967: 366-369).

zung zu finden ist, nach welcher der Autor den Terminus verwendet. Der Autor lehnt sich in seiner Sprachverwendung also an den Sprachgebrauch eines anderen Autors oder einer bestimmten Fachdomäne an, ohne die dort gültige Definition in seinem Text zu wiederholen.

Beispiel:

Ich werde den Ausdruck *Annotation* in dem in der Hypertextliteratur eingebürgerten und (Tochtermann 1995) entsprechenden Sinn verwenden.

Für die manuelle Annotation von definitorischen Textsegmenten in den vier „Pilottexten“ wurde eine Dokumentgrammatik konzipiert, die folgende Anforderungen erfüllen sollte:

- (a) sie sollte in kategorialer Hinsicht so granular sein, dass innerhalb der in den zugehörigen Dokument-Instanzen ausgezeichneten Textsegmente die basalen Konstituenten von Definitionen – *Definiendum* und *Definiens* – (soweit möglich) als diskrete Einheiten herausgegriffen werden können (dies mit Blick auf eine Sekundärverwertung der Dokument-Instanzen für eine automatische Generierung von Glossaren oder lexikographischen/terminologischen Datenbanken);
- (b) sie sollte im Falle des Eingebettetseins einer Definition in eine für das Verständnis notwendige größere syntaktische oder textuelle Struktur das Herausgreifen nicht lediglich der Definition als solcher, sondern ggf. auch des zugehörigen verständnisrelevanten Kontexts erlauben;
- (c) sie sollte in Fällen einer elliptischen Angabe von Mehrwort-Termini in Definiendum-Position die Möglichkeit der Angabe der entsprechenden Vollform erlauben (dies mit Blick auf die fehlerfreie automatische Zuordnung einer Definition zu einem Terminus);
- (d) sie sollte in Anbetracht der vielfältigen Versprachlichungsmöglichkeiten von Definitionen so flexibel sein, dass sich mit einer möglichst einfachen Modellierungsstruktur möglichst alle Versprachlichungsvarianten erfassen lassen.

Darüber hinaus sollte aus verarbeitungspraktischen Gründen die Dokumentgrammatik zugleich die in Schritt 2 (vgl. 2.2) vorzunehmende automatische Annotation von Termverwendungsinstanzen mit abbilden, um Definitionen und Termverwendungsinstanzen auf ein- und derselben Annotationsebene behandeln zu können.

Als Repräsentationsformat für die Dokumentgrammatik wurde eine XML-DTD gewählt. Um Punkt (d) der o.a. Anforderungen gerecht zu werden, wurden für die zentralen Elemente der DTD Inhaltsmodelle definiert, die hinsichtlich der Abfolge und Auftretenshäufigkeit der einzelnen Kindelemente weitgehend variierbar sind. Der hierdurch bewusst in Kauf genommene (und der Bandbreite an Vertextungsmöglichkeiten geschuldete) Verlust an Restriktivität wurde dadurch ausgeglichen, dass zu den einzelnen in der DTD beschriebenen Elementen Richtlinien für die Annotation formuliert wurden. Diese besagten beispielsweise, dass eine Definition notwendigerweise *mindestens ein* definiendum-Element sowie entweder *genau ein* definiens-Element oder aber (im Falle im Text diskontinuierlich auftretender Definienses) *mindestens zwei* definiens-Segmente enthalten muss.

Nachfolgend die kommentierte DTD für die Annotation von definitorischen Textsegmenten sowie einige Annotationsbeispiele aus den Dokument-Instanzen:

```
<!--  
Projekt HyTex, DFG-Forschergruppe "Texttechnologische Informationsmodellierung"  
http://www.hytex.info/
```

```

=====
Kommentierte DTD zur Annotation von definitorischen
Textsegmenten und Termverwendungsinstanzen
(v. 2.0)
=====
-->

<!--
Die DTD ist so konzipiert, dass Termverwendungsinstanzen und Definitionen
in verschiedenen Annotationsprozessen getrennt voneinander annotiert werden
koennen. Die DTD kann auch zur Validierung nur eines der beiden verwendet
werden.

An vielen Stellen war es daher und aufgrund der Beschraenkungen von DTDs
(insbesondere im Fall gemischter Inhaltsmodelle) notwendig, die Inhaltsmo-
delle sehr allgemein zu fassen. Wo inhaltliche Beschraenkungen notwendig
sind, die nicht in der DTD ausgedrueckt werden koennen, wurde dies dokumen-
tiert.
-->

<!--
=====
ELEMENT: Wurzelement d e f i n i t i o n s
=====
-->

<!ELEMENT definitions (#PCDATA | defSegment | term)*>

<!-- Das Wurzelement d e f i n i t i o n s beschreibt ein Dokument, in
welchem Termverwendungsinstanzen (TVIs) und/oder defitinerische Textsegmen-
te asgezeichnet wurden. Durch das sehr allgemein gehaltene Inhaltsmodell
kann die Struktur entweder nur fuer die Auszeichnung definitorischer Text-
segmente oder nur fuer die (automatische) Auszeichnung von TVIs oder fuer
beides (und fuer die anschliessende Validierung der Instanzen) verwendet
werden.
-->

<!--
=====
ELEMENT: d e f S e g m e n t
=====
-->

<!ELEMENT defSegment (#PCDATA | def | defComplex | term)*>

<!-- Das Element d e f S e g m e n t beschreibt solche Textsegmente, in de-
nnen mindestens eine Definition oder ein Definitionskomplex enthalten ist. d
e f S e g m e n t kann jedoch gegenueber Instanzen der Elemente d e f (fuer
Definitionen) bzw. d e f C o m p l e x zudem noch textuellen Kontext umfas-
sen. Impetus bei der Beschreibung einzelner Textteile als d e f S e g m e n
ts ist es, fuer eine moegliche Anwendung textuelle Einheiten herausgreifen
zu koennen, die folgende Kriterien erfuellen:
(a) Sie sind mindestens syntaktisch abgeschlossen (d.h.: repraesentieren
entweder einen ganzen Satz oder ein Syntagma, das auch ohne Kontext autonom
verstaendlich ist und somit fuer die Konzeptualisierung des darin definier-
ten Terminus hilfreich sein kann),
(b) Sie sind idealitaer auch kohaesiv geschlossen (d.h.: in Faellen, in
welchen die definiens- oder die definidendum-Position mit einer Anapher be-
setzt ist, umfasst das betreffende d e f S e g m e n t nach Moeglichkeit
auch noch den zur Kohaerenzbildung noetigen Kontext - ein d e f S e g m e n
t sollte jedoch (pi mal Daumen) maximal drei Saetze umfassen, damit der in
der Anwendung angebotene Textausschnitt kompakt bleibt.

```

Das Element `d e f S e g m e n t` bildet das Container-Element fuer alle Definitionen und Definitionskomplexe.

-->

<!--

=====

ELEMENT: `d e f`

=====

-->

**<!ELEMENT def (#PCDATA | term | definiendum |
definiens | definiensSegment)*>**

<!ATTLIST def

**type (Setzung | Selbstzuschreibung |
 Direktive | Fremdzuschreibung) #REQUIRED>**

<!-- Das Element `d e f` beschreibt einzelne definitorische Textsegmente. Einzelnen Instanzen des elements `d e f` wird jeweils (obligatorisch) ein `t y p e` zugewiesen, dessen zulaessiger Wertebereich die Bezeichnungen der im Projekt zugrundegelegten Sprachhandlungstypen umfasst. Die `t y p e` zugewiesenen Werte sind zentral fuer die spaetere automatische Gewichtung von Definitionen im Falle von "Definitionen-Konkurrenz" in ein- und demselben Text.

Aufgrund oder Variabilitaet bei der Vertextung von Definitionen ist das Inhaltsmodell von `d e f` recht allgemein gehalten: Sowohl ist die Reihenfolge des Auftretens der Kindelemente `d e f i n i e n d u m` und `d e f i n i e n s` beliebig, als auch kann das `d e f i n i e n s` diskontinuierlich auftreten (in Form zweier oder mehrerer Vorkommnisse des Elements `d e f i n i e n s S e g m e n t`). Weiterhin kann das Element `d e f` auch PCDATA beinhalten (naemlich solche Textteile, die nicht zur eigentlichen `d e f`-Struktur gehoeren, aber die z.B. zwischen zwei Vorkommnissen von `d e f i n i e n s S e g m e n t` stehen).

Beispiele fuer die Variation bei der Vertextung von Definitionen:

Es sind vier Strukturtypen von definitorischen Textsegmenten moeglich, die sich in der Abfolge von `Definiens` und `Definiendum` unterscheiden:

[[[Legende: `*...*` markiert in den nachfolgenden Beispielen das `definiendum`, `#...#` markiert das `Definiens` bzw. `DefiniensSegmente`]]]

(1) `definiendum` vor `definiens`

Bsp.: `*Links*` sind `#computerverwaltete Zuordnungen zwischen Ankern#`.

(2) `definiendum` vor diskontinuierlichem (d.h.: durch anderen Text durchbrochenem) `definiens`

Bsp.: Unter `*Annotationen*` werden in der Hypertextliteratur `#Anmerkungen und Notizen#` verstanden, `#die ein Hypertextnutzer waehrend des Rezeptionsvorgangs zu den Inhalten eines Moduls anbringt#`.

(3) `definiens` vor `definiendum`

Bsp.: `#Computerverwaltete Zuordnungen zwischen Ankern#` bezeichnet man als `*Links*`.

(4) diskontinuierliches `definiens` umschliesst das `definiendum`:

Bsp.: In der Hypertextliteratur werden solche `#Anmerkungen und Notizen#` als `*Annotationen*` bezeichnet, `#die ein Hypertextnutzer waehrend des Rezeptionsvorgangs zu den Inhalten eines Moduls anbringt#`.

Richtlinien fuer die Annotation:

d e f enthaelt

- (1) mindestens ein d e f i n i e n d u m-Element sowie
- (2) entweder genau ein d e f i n i e n s-Element oder mindestens zwei d e f i n i e n s S e g m e n t-Elemente.

Prinzipiell kann das Element d e f zwei Vorkommnisse des Kindelements d e f i n i e n d u m beinhalten, naemlich in solchen Faellen, in welchen Benennungsalternativen angegeben werden; Beispiel:

#Filter, die ueber etadaten zu Modulen und Links operieren#, nenne ich im Weiteren *metadatenorientierte Filter*, kurz: *Metafilter*.

-->

<!--

=====

ELEMENT: d e f C o m p l e x

=====

-->

<!ELEMENT defComplex (#PCDATA | term | definiendum | definiensAlt)*>

<!ATTLIST defComplex

 type (Setzung | Selbstzuschreibung |
 Direktive | Fremdzuschreibung) #REQUIRED >

<!-- Erlaeuterung zum Element d e f C o m p l e x: Ein Definitions-Komplex liegt dann vor, wenn in eine syntaktische Struktur zwei alternative Definiertes zu ein und demselben Terminus eingebettet sind und der Terminus (als Definiendum) dabei nur einmal genannt wird.

Beispiel:

Hypertext-Module werden in der Werkstattsprache des WWW oft als "Seiten" bezeichnet, wobei "*Seite*" einmal #ein am Bildschirm sichtbares Objekt# (Seite-Screen) ein anderes Mal #die von einem WWW-Server verwaltete Dateneinheit# (Seite-Datei) benennt.

Richtlinien fuer die Annotation:

d e f C o m p l e x enthaelt

- (1) mindestens ein d e f i n i e n d u m-Element sowie
- (2) mindestens zwei d e f i n i e n s A l t-Elemente.

-->

<!--

=====

ELEMENT: d e f i n i e n d u m

=====

-->

<!ELEMENT definiendum (#PCDATA | term)*>

<!ATTLIST definiendum

 baseForm CDATA #IMPLIED>

<!-- Eine Instanz des Elements d e f i n i e n d u m enthaelt nach dem automatischen Annotierungsprozess der Termverwendungsinstanzen normalerweise genau ein t e r m-Element. In diesem Fall wird beim Linking das Attribut b a s e F o r m von d e f i n i e n d u m nicht ausgewertet, sondern nur das von t e r m. In Faellen, in denen das Definiendum jedoch keinen Terminus aus der Termlist enthaelt, da dieser im Definitionstext nur implizit vorkommt, enthaelt d e f i n i e n d u m nach der automatischen Annotierung kein t e r m-Element. Fuer diesen Fall muss das Attribut b a s e F o r m von d e f i n i e n d u m bei der Annotation der Definitionen manuell ge-

setzt werden, um auch diese Termverwendungen spaeter automatisch auffinden zu koennen.

Beispiel:

```
<defSegment>
Hyperlinks koennen nach ihrem Zielpunkt kategorisiert werden, je nach
dem, ob <def type="Fremdzuschreibung"> <definiens>eine Stelle in der
gleichen informationellen Einheit</definiens> &ndash;&gt;
(<definiendum baseForm="intrahypertextueller
Link">intrahypertextuell</definiendum>)
</def>, <def type="Fremdzuschreibung"> <definiens>eine Informationseinheit
in der
gleichen Hypertextbasis</definiens>
(<definiendum baseForm="interhypertextueller
Link">interhypertextuell</definiendum>)
</def> oder <def type="Fremdzuschreibung"> <definiens>in einer anderen
Hypertextbasis</definiens>
(<definiendum baseForm="extrahypertextueller
Link">extrahypertextuell</definiendum>)
</def> referenziert wird (Kuhlen 1991: 107f.; Kennedy/Musciano 1999:
213;215ff.).
</defSegment>
```

Richtlinien fuer die Annotation:

* enthaelt null bis ein t e r m-Elemente

-->

<!--

=====

ELEMENT: d e f i n i e n s

=====

-->

<!ELEMENT definiens (#PCDATA | term)*>

<!--

=====

ELEMENT: d e f i n i e n s S e g m e n t

=====

-->

<!ELEMENT definiensSegment (#PCDATA | term)*>

<!--

=====

ELEMENT: d e f i n i e n s A l t

=====

-->

<!ELEMENT definiensAlt (#PCDATA | term)* >

<!--

=====

ELEMENT: t e r m

=====

Vorkommnisse des Elements t e r m werden auf der Grundlage der Termlist au-
tomatisch annotiert.

-->

<!ELEMENT term (#PCDATA)>

<!ATTLIST term

baseForm CDATA #REQUIRED

Annotationsbeispiele aus den Dokument-Instanzen:

```
<defSegment>
  <def type="Selbstzuschreibung">
    Als
    <definiendum>E-Texte</definiendum>
    bezeichne ich
    <definiens>linear organisierte Texte, die in ein
    Hypernetz eingebunden sind</definiens>
  .
  </def>
</defSegment>

<defSegment>
  <def type="Setzung">
    <definiendum>Hypertextdokumente</definiendum>
    -- das steckt als Idee hinter der vorgeschlagenen
    Differenzierung -- sind
    <definiens>Hyperdokumente mit Textstatus, die aber im Ge-
    gensatz zu den E-Texten nicht-linear organisiert
    sind</definiens>
  .
  </def>
</defSegment>

<defSegment>
  So bezeichnet der Ausdruck "Link" einerseits
  <def type="Setzung">
    <definiens>die auf der Nutzeroberfläche wahrnehmbaren
    Bildschirmobjekte</definiens>
    (
    <definiendum>die Linkanzeiger</definiendum>
    )
  </def>
  ,
  andererseits
  <def type="Setzung">
    <definiens>die auf der konzeptionellen Ebene
    modellierten Hypertextobjekte</definiens>
    (
    <definiendum>die Links</definiendum>
    )
  </def>
  .
</defSegment>

<defSegment>
  Zu einem Hypertextsystem gehört
  <def type="Fremdzuschreibung">
    <definiens>Software, die den Aufbau, die Bearbeitung und
    Veränderung von Hypertextbasen unterstützt</definiens>
    ; diese wird häufig als
    <definiendum>Autorenwerkzeug</definiendum>
    („
    <definiendum>Authoring Tool</definiendum>
```

```

        „)
        bezeichnet
    </def>
</defSegment>

<defSegment>
    <defComplex type="Fremdzuschreibung">
        <definiendum>Der Ausdruck "Hypertext"</definiendum>
        wird in der Literatur zur
        <definiensAlt>Bezeichnung einer neuen Schreib- und
        Lesetechnologie</definiensAlt>
        erstens, der
        <definiensAlt>dafür verwendeten Software
        (dem Hypertextsystem)</definiensAlt>
        zweitens und
        <definiensAlt>der von dieser erstellten verwalteten
        Dokumente</definiensAlt>
        drittens verwendet.
    </defComplex>
</defSegment>

```

2.2 Schritt 2: Erzeugung einer Termliste und automatische Annotation von Termverwendungsinstanzen

Sämtliche Termini, die in den in Schritt 1 annotierten Definitionen in Definiendum-Position auftraten wurden in Schritt 2 in eine Liste definierter Termini überführt. Für die einzelnen Einträge in dieser Liste wurde eine Grundform (die lexikographische „Nennform“) festgelegt und anhand einer Anzahl an „derived forms“ eine Auflistung sämtlicher flexionsmorphologischer Varianten angegeben. Auf Grundlage der somit erweiterten Liste wurden dann automatisch sämtliche Verwendungsinstanzen der Termini in den vier „Pilottexten“ annotiert. Für diesen Zweck wurde der Termverwendungsinstanzen-Tagger *tvi-tagger.pl* (kurz „Tiffi“) entwickelt⁵, der anhand der Wortformenliste jede Termverwendungsinstanz in einer XML- oder Textdatei mit <term>-Tags umschließt beziehungsweise die <term>-Tags aktualisiert (d.h.: ggf. hinzufügt oder löscht, sofern in der betreffenden Datei bereits <term>-Tags gesetzt wurden).

Die Grundlage für die XML-Repräsentation der Termini und zugehörigen Flexionsformen („Wortformenliste“) bildete folgende DTD:

```

<!ELEMENT wordformlist (term*) >
<!ELEMENT term (baseForm, derivedForm*) >
<!ATTLIST term
  normalForm CDATA #REQUIRED
  id ID #REQUIRED >
<!ELEMENT baseForm (#PCDATA) >
<!ELEMENT derivedForm (#PCDATA) >

```

5 Programmierer: Jan Frederik Maas, Programmiersprache: Perl.

Auszug aus der entsprechend der DTD modellierten Liste der Termini und zugehörigen Flexionsformen:

```
<term id="id29" normalForm="Dokument, wohlgeformtes">
    <baseForm>wohlgeformtes Dokument</baseForm>
    <derivedForm>wohlgeformten Dokuments</derivedForm>
    <derivedForm>wohlgeformten Dokument</derivedForm>
    <derivedForm>wohlgeformte Dokument</derivedForm>
    <derivedForm>wohlgeformte Dokumente</derivedForm>
    <derivedForm>wohlgeformten Dokumente</derivedForm>
    <derivedForm>wohlgeformten Dokumenten</derivedForm>
</term>
<term id="id31" normalForm="Dokument-Instanz">
    <baseForm>Dokument-Instanz</baseForm>
    <derivedForm>Dokument-Instanzen</derivedForm>
</term>
<term id="id90" normalForm="Link, intensional definierter">
    <baseForm>intensional definierter Link</baseForm>
    <derivedForm>intensional definierter Links</derivedForm>
    <derivedForm>intensional definierte Links</derivedForm>
    <derivedForm>intensional definierten Link</derivedForm>
    <derivedForm>intensional definierten Links</derivedForm>
    <derivedForm>intensional definierten Linkes</derivedForm>
    <derivedForm>intensional definiertem Link</derivedForm>
</term>
<term id="id172" normalForm="Traversierung">
    <baseForm>Traversierung</baseForm>
    <derivedForm>Traversierungen</derivedForm>
</term>
<term id="id173" normalForm="Traversierungsattribut">
    <baseForm>Traversierungsattribut</baseForm>
    <derivedForm>Traversierungsattributs</derivedForm>
    <derivedForm>Traversierungsattribute</derivedForm>
    <derivedForm>Traversierungsattributen</derivedForm>
    <derivedForm>Traversierungsattributes</derivedForm>
</term>
```

2.3 Schritt 3: Typisierung von definitorischen Textsegmenten auf Basis einer pragmatischen Typologie definitorischen Sprachhandelns

Ziel von Schritt 3 war es, die in Schritt 1 ausgezeichneten definitorischen Textsegmente zu typisieren, um mit Blick auf das gewählte Anwendungsszenario eine Möglichkeit zu schaffen, im Falle mehrerer in einem Fachtext miteinander „konkurrierender“ Definitionen zu ein- und demselben Terminus diese Definitionen auf der Grundlage einer Hierarchie von Typen hinsichtlich ihrer Relevanz für die Termverwendung des betreffenden Fachtextautors gegeneinander gewichten zu können.

Zentral für die Frage, welche von mehreren zur Verfügung stehenden Definitionen diejenige ist, die für die Verwendung eines Terminus durch den Autor für verbindlich erachtet werden kann, sind pragmatische Überlegungen: Wenn ein Autor einerseits in seinem Text mehrere Definitionen zu ein- und demselben Terminus gibt und wir ihm andererseits unterstellen, dass seiner eigenen (Fach-)Sprachverwendung ein widerspruchsfreies Terminologiesystem unterliegt, dann muss die Tatsache, dass er mehrere Definitionen zu ein- und demselben Terminus gibt, auf bestimmte Strategien zurückzuführen sein. Nicht selten finden sich in Fachtexten

Abschnitte, in denen es dem jeweiligen Autor um die Klärung der für den behandelten Gegenstand zentralen Konzepte und die für die lexikalische Kartographierung des Konzeptbereichs zu verwendenden Termini zu tun ist. In solchen Abschnitten werden häufig verschiedene Konzeptualisierungen zu ein- und demselben terminologischen Ausdruck oder aber verschiedene Benennungen zu einem Konzept angeführt, diskutiert und gegeneinander abgewogen, bis sich der Autor schließlich für eine der referierten Alternativen entscheidet oder aber für seine eigene Sprachverwendung im Folgetext eine explizite Festlegung vornimmt (z.B. der Form „Ich hingegen möchte im Folgenden unter einem X ein Y verstehen“). Den verschiedenen Textsegmenten, in welchen vom Autor im Zuge solcher „Terminologiediskussionen“ jeweils unterschiedliche Definitionen angegeben werden, unterliegen dabei einerseits unterschiedliche Handlungszwecke, andererseits werden in propositionaler Hinsicht z.T. auch unterschiedliche Gültigkeitsansprüche an sie geknüpft.

Wenn ein Autor schreibt:

Mesonen sind zusammengesetzte Teilchen, die aus je einem Quark und einem Antiquark bestehen,

so trifft er damit eine Feststellung über das Zutreffen des Etiketts *Mesonen* auf Klassen von Objekten mit dem Merkmal „Teilchen, die aus je einem Quark und einem Antiquark bestehen“ und unterlegt diese Feststellung zugleich mit einem universalen Gültigkeitsanspruch (gibt also implizit zu erkennen, dass er selbst dieser Definition zustimmt).

Wenn hingegen ein Autor (der selbst nicht Physiker ist) schreibt:

Unter Mesonen versteht man in der Physik zusammengesetzte Teilchen, die aus je einem Quark und einem Antiquark bestehen,

so trifft er damit eine Feststellung über den Gebrauch des Ausdrucks *Mesonen* in der Fachsprache der Physik und beansprucht zugleich, dass das Zutreffen des Etiketts *Mesonen* auf Klassen von Objekten mit dem Merkmal „Teilchen, die aus je einem Quark und einem Antiquark bestehen“ in der Fachsprache der Physik gültig ist. Er beansprucht damit jedoch *nicht*, dass die somit gegebene Definition von *Mesonen* auch für seine eigene Verwendung des Ausdrucks *Mesonen* zutrifft (es sei denn, er macht im weiteren Text explizit deutlich, dass er sich dieser in der Physik etablierten Definition anschließen möchte).

Wenn jedoch ein Autor schreibt:

Unter Mesonen verstehe ich zusammengesetzte Teilchen, die aus je einem Quark und einem Antiquark bestehen,

dann trifft er keine *Feststellung*, sondern nimmt eine *Festsetzung* vor, d.h.: er legt für seine eigene Sprachverwendung einen ganz bestimmten Gebrauch des Ausdrucks *Mesonen* fest und geht somit gegenüber den Adressaten seines Textes die Verbindlichkeit ein, wenn er fortan den Ausdruck *Mesonen* verwendet, dies im Sinne der somit vorgenommenen Festsetzung zu tun. Der Gültigkeitsanspruch einer solchen Aussage wird also explizit als auf die eigene Sprachverwendung eingeschränkt deklariert. Aussagen dieser Art sind somit (im Gegensatz zu den beiden erstgenannten) nur bedingt falsifizierbar – bestenfalls kann man dem Autor nachweisen, dass es nicht zutrifft, dass er sich im weiteren Verlauf seines Textes an diese somit gesetzte „Sprachverwendungsverpflichtung“ hält oder aber man kann ihm (was dann natürlich nichts mehr mit einer Bewertung im Sinne der WAHR/FALSCH-Dimension zu tun hat) eine Unzweckmäßigkeit seiner Definition vorwerfen:

Feststellungen über die Sprache sind empirische Behauptungen über eine Wissenschaftssprache; daher kann man mit ihnen recht oder unrecht haben wie sonst mit empirischen Behauptungen auch. [...] Festsetzungen können nicht in diesem Sinne falsch sein. [...] Man kann auch Festsetzungen angreifen, nur eben auf andere Weise als Feststellungen. Festsetzungen für die Sprache trifft man zu praktischen Zwecken: um es sehr allgemein auszudrücken – zum Zwecke besserer Verständigung. Festsetzungen können daher kritisiert werden je nachdem, wie gut sie diesen Zweck erreichen. (Savigny 1970: 23f.)

In jedem Falle sind definitorischen Aussagen letzterer Art eher *performativ*, wohingegen die beiden ersteren eher als *konstativ* zu erachten sind.

In der „klassischen“ Definitionstheorie gibt es die Unterscheidung zwischen *Feststellungsdefinitionen* und *Festsetzungsdefinitionen* (siehe z.B. Savigny 1970). Für das im HyTex-Projekt benötigte Kategorieninventar zur Typisierung von Definitionen wurde zunächst von dieser Unterscheidung ausgegangen, da „Feststellung“ und „Festsetzung“ Kategorien sind, die sich pragmatisch begründen lassen, insofern sie Auskunft über die einer definitorischen Äußerung unterliegende Handlungsabsicht des Äußernden geben:

Festsetzende Definitionen sind keine Aussagen und können daher nicht wahr oder falsch sein. Als Sprechakte betrachtet reichen sie von Willensbekundungen (z.B. in einem Vortrag ein bestimmtes Wort stets in einem bestimmten Sinne zu gebrauchen) und Selbstverpflichtungen – soweit der private Sprachgebrauch betroffen ist – über Aufforderungen und Empfehlungen bis zu verbindlichen Wortverwendungsnormen (z.B. in Form juristischer Definitionen) – soweit der öffentliche Sprachgebrauch betroffen ist. (Enzyklopädie Philosophie und Wissenschaftstheorie: 439)

Unter Rückgriff auf den funktionalpragmatischen Ansatz der „Grammatik der deutschen Sprache“ (GDS) wurde die mit „Feststellung“ und „Festsetzung“ eröffnete Kategorisierungsmöglichkeit auf eine dezidiert pragmatische Basis gestellt, was die Möglichkeit einer differenzierten Typologie eröffnete, die Definitionen auf erster Ebene nach Zweckbereichen sprachlichen Handelns, auf zweiter Ebene nach den jeweils gewählten Sprachhandlungstypen und auf dritter Ebene unter propositionalen Gesichtspunkten einteilt.

Definitorische Äußerungen sind nach unserem Entwurf *Sprachhandlungszüge*, die auf einem *definitorischen Sachverhaltsentwurf* basieren, der (a) zur Erreichung eines bestimmten *kommunikativen Zwecks* hervorgebracht wird, für den (b) ein bestimmter Gültigkeitsanspruch erhoben wird, und (c) dessen Versprachlichung an einem bestimmten *Versprachlichungsmuster* orientiert ist. Definitorische Äußerungen werden somit sowohl unter pragmatischem (a) und propositionalem Aspekt (b) betrachtet und lassen sich darüber hinaus auch hinsichtlich der gewählten sprachlichen Realisierungsform (c) beschreiben. Für die Typisierung definitorischer Textsegmente spielt der Versprachlichungsaspekt (c) keine direkte Rolle, insofern die Wahl eines bestimmten Versprachlichungsmusters (z.B. die Entscheidung für eine Nominal- anstatt eine Realdefinition) nicht notwendigerweise an den jeweils gewählten Handlungstyp geknüpft sein muss. Beispielsweise können bestimmte „Festsetzungsdefinitionen“ sowohl in Form von Nominaldefinitionen als auch (wenn auch eher selten) in Form von Realdefinitionen realisiert werden:

Unter Mesonen verstehe ich zusammengesetzte Teilchen, die aus je einem Quark und einem Antiquark bestehen [Nominaldefinition; sprachbezogen]

Mesonen sind für mich zusammengesetzte Teilchen, die aus je einem Quark und einem Antiquark bestehen [Realdefinition; Sachaussage]

Der Handlungsaspekt (a) sowie der mit einer Proposition verknüpfte Gültigkeitsanspruch (b) sind jedoch für ein angestrebtes „Ranking“ von konkurrierenden Definitionen überaus relevante Faktoren.

Auf erster Ebene teilen wir definitorische Äußerungen nach Zweckbereichen sprachlichen Handelns in solche Äußerungen ein, die auf *Handlungskoordination* gerichtet sind und in solche, die auf einen *Transfer von Wissen* gerichtet sind. Dies entspricht weitgehend der Einteilung *Festsetzungsdefinition – Feststellungsdefinition*: Auf Handlungskoordination gerichtete definitorische Äußerungen zielen darauf, explizit Verbindlichkeiten in Bezug auf die weitere Sprachverwendung festzulegen, oder – wie Savigny (1970: 24) es für Festsetzungsdefinitionen formuliert – „zu praktischen Zwecken“, „zum Zwecke besserer Verständigung“. Auf Wissenstransfer gerichtete definitorische Äußerungen zielen darauf, eine Feststellung über einen definitorischen Sachverhalt (ggf. eingeschränkt auf die Weltsicht bzw. den Sprachgebrauch eines bestimmten anderen Autors, einer bestimmten anderen Schule oder einer bestimmten Fachdomäne) zu treffen.

Definitorische Äußerungen im Zweckbereich *Wissenstransfer* sind immer *assertiv*, d.h.: mit ihnen wird eine falsifizierbare Behauptung aufgestellt. Für definitorische Äußerungen im Zweckbereich *Handlungskoordination* lassen sich hingegen zwei Handlungstypen unterscheiden, nämlich *kommissiv* und *direktiv* gemeinte Handlungen. *Kommissiv* gemeinte definitorische Sprachhandlungen dienen einem Autor dazu, seinem Adressaten gegenüber Festlegungen in bezug auf das eigene (zukünftige) Sprachhandeln zu treffen; mit *direktiv* gemeinten definitorischen Sprachhandlungen geht der Autor solche Verpflichtungen zwar ebenfalls ein – nur sind sie im Gegensatz zu kommissiven Äußerungen zudem auch auf das Sprachhandeln der Adressaten und damit auf eine Veränderung des Sprachgebrauchs in der betreffenden Fachdomäne gerichtet. Kommissive definitorische Sprachhandlungen bezeichnen wir als *definitorische Selbstzuschreibungen*, direktive definitorische Sprachhandlungen als *definitorische Direktiven*.

Da auf Handlungskoordination gerichtete definitorische Äußerungen per se nicht falsifizierbar sind (s.o.), lassen sie sich unter propositionalem Aspekt nicht weiter unterteilen. Assertive im Bereich Wissenstransfer hingegen lassen sich in einem weiteren Typologisierungsschritt danach unterteilen, wie universal der Gültigkeitsanspruch ist, der für den proponierten definitorischen Sachverhalt erhoben wird. Dieser kann entweder universal sein („definitorische Behauptungen / Setzungen“⁶) oder in spezifischer Weise eingeschränkt („definitorische Fremdzuschreibungen“). Bei definitorischen Fremdzuschreibungen ist aufgrund expliziter Angabe eines Sprachverwenders oder einer Sprachverwendergruppe die Gültigkeit des definitorischen Sachverhalts immer auf bestimmte Kontexte bzw. Weltsichten eingeschränkt.

6 Im Unterschied zur mit dem Begriffspaar *Festsetzungsdefinition – Feststellungsdefinition* verbundenen Benennungsmotivik wurde in unserem Typologieansatz der Ausdruck *Setzung* reserviert für assertive Aussagen (Behauptungen) mit universalem Gültigkeitsanspruch: Mit definitorischen Aussagen wie „Ein X ist ein Y“ wird einschränkungslos eine Zuordnung $X = Y$ gesetzt (behauptet), die entweder wahr oder falsch sein kann.

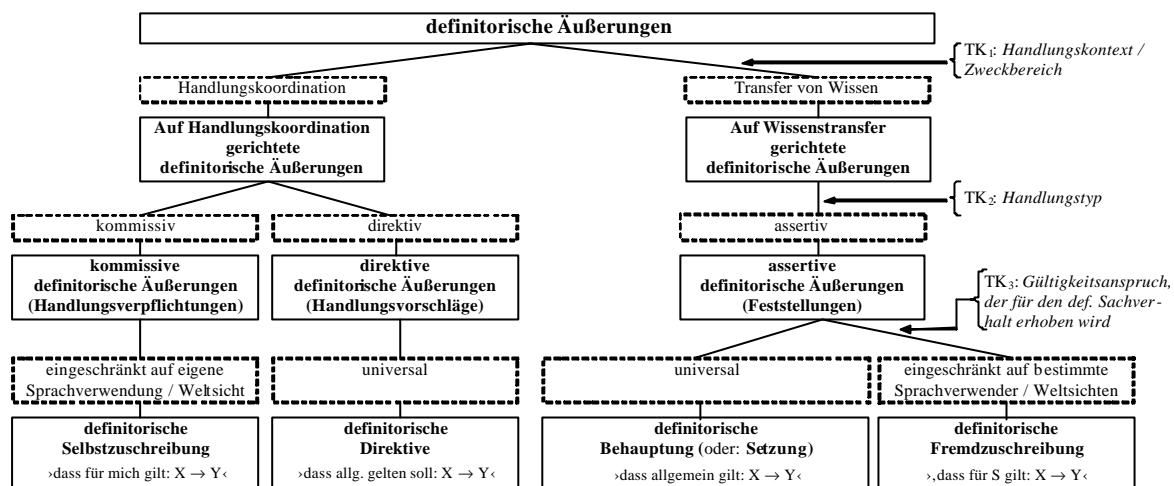


Abb. 2: Pragmatisch motivierte Typologie definitorischer Äußerungen.

Beispiele:

Unter Morphemen verstehe ich im Folgenden die kleinsten bedeutungstragenden Einheiten einer Sprache. [*Selbstzuschreibung*]

Ich schlage vor, die kleinsten bedeutungstragenden Einheiten einer Sprache als Morphem zu bezeichnen. / Unter einem Morphem hat man die kleinste bedeutungsunterscheidende Einheit einer Sprache zu verstehen. [*Direktive*]

Morpheme sind die kleinsten bedeutungsunterscheidenden Einheiten einer Sprache. [*Behauptung, „Setzung“*]

Die strukturalistische Linguistik versteht unter einem Morphem die kleinste bedeutungsunterscheidende Einheit einer Sprache. [*Fremdzuschreibung*]

2.4 Schritt 4: Erarbeitung von Ranking-Regeln für das automatische Linking

Um einem 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, wird die pragmatische Typisierung definitorischer Textsegmente dazu genutzt, die automatische 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). Die unterschiedlichen Grade an Verbindlichkeit, die ein Autor mit der Wahl eines der in 2.3 beschriebenen Typen für sein Sprachhandeln eingeht, sowie die unterschiedlichen Gültigkeitsansprüche, die assertiven Sprachhandlungen auf propositionaler Ebene unterlegt werden können, sprechen für die Annahme der folgenden Grundregeln für eine Gewichtung von Definitionen⁷:

7 *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 Vor- text einer Verwendungsinstanz von X höhere Priorität als B“.

(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 Spezifizierungsgrad 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) 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 Definiertes charakterisierten Konzepte miteinander kompatibel sind. In Fällen, in welchen zwei Fremdzuschreibungen miteinander konkurrieren (ohne dass zum selben Terminus eine Definition höherrangigen Typs vorliegt), nehmen wir an, dass sich bei einer der beiden die Angabe der Sprachverwendergruppe bzw. Weltsicht, auf welche die Gültigkeit des definitiven Sachverhalts eingeschränkt ist, so interpretieren lässt, dass der Autor der betreffenden Sprachverwendergruppe bzw. Weltsicht zugerechnet werden kann – etwa in folgendem Beispiel einer Konkurrenz zweier definitiver 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 oben angeführten Grundregeln lassen sich wie folgt ausformulieren:⁸

- | | |
|-----------------|---|
| Regel 1: | <i>SelbstZ</i> schlägt <i>FremdZ</i> |
| Regel 2: | <i>Dir</i> schlägt <i>FremdZ</i> |
| Regel 3: | <i>Setzg</i> schlägt <i>FremdZ</i> |
| Regel 4: | Konkurriert <i>SelbstZ</i> mit <i>Dir</i> , dann sind sie kompatibel. |
| Regel 5: | Konkurriert <i>SelbstZ</i> mit <i>Setzg</i> , dann sind sie kompatibel. |
| Regel 6: | Konkurriert <i>Dir</i> mit <i>Setzg</i> , dann sind sie kompatibel. |
| Regel 7: | Zwei konkurrierende <i>SelbstZ</i> sind miteinander kompatibel. |
| Regel 8: | Zwei konkurrierende <i>Dir</i> sind miteinander kompatibel. |
| Regel 9: | Zwei konkurrierende <i>Setzg</i> sind miteinander kompatibel. |

Diese Grundregeln beziehen sich zunächst ausschließlich auf Fälle, in welchen zwei Definitionen mit einander konkurrieren. Auch ist in ihnen die Reihenfolge des Auftretens der einzel-

8 Legende: *SelbstZ* = Selbstzuschreibung, *FremdZ* = Fremdzuschreibung, *Dir* = Direktive, *Setzg* = Setzung.

nen Definitionen im Textverlauf (relativ zu einer je konkreten Termverwendungsinstanz) noch nicht berücksichtigt. Die Grundregeln wurden daher unter Einbeziehung des Kriteriums ihrer Auftretensreihenfolge weiter ausdifferenziert.

Regel Nr.	Der Verwendungsinstanz eines Terminus X gehen zwei Definitionen zu X voraus, und zwar in folgender Reihenfolge:		„Gewinner“	Kommentar
	Def-Typ in Position 1⁹	Def-Typ in Position 2		
1.1	<i>SelbstZ</i>	<i>FremdZ</i>	<i>SelbstZ</i>	
1.2	<i>FremdZ</i>	<i>SelbstZ</i>	<i>SelbstZ</i>	
2.1	<i>Dir</i>	<i>FremdZ</i>	<i>Dir</i>	
2.2	<i>FremdZ</i>	<i>Dir</i>	<i>Dir</i>	
3.1	<i>Setzg</i>	<i>FremdZ</i>	<i>Setzg</i>	
3.2	<i>FremdZ</i>	<i>Setzg</i>	<i>Setzg</i>	
4.1	<i>SelbstZ</i>	<i>Dir</i>	<i>Dir</i>	Annahme: Die <i>SelbstZ</i> und die <i>Dir</i> sind miteinander kompatibel. Möglicherweise stellt die <i>Dir</i> eine Präzisierung der <i>SelbstZ</i> dar. Daher gewinnt in diesem Fall die <i>Dir</i> .
4.2	<i>Dir</i>	<i>SelbstZ</i>	<i>SelbstZ</i>	Annahme: Die <i>Dir</i> und die <i>SelbstZ</i> sind miteinander kompatibel. Möglicherweise stellt die <i>SelbstZ</i> eine Präzisierung der <i>Dir</i> dar. Daher gewinnt in diesem Fall die <i>SelbstZ</i> .
5.1	<i>SelbstZ</i>	<i>Setzg</i>	<i>SelbstZ</i>	Annahme: Die <i>SelbstZ</i> und die <i>Setzg</i> sind miteinander kompatibel. Aufgrund des kommissiven Charakters der <i>SelbstZ</i> gewinnt aber die <i>SelbstZ</i> .
5.2	<i>Setzg</i>	<i>SelbstZ</i>	<i>SelbstZ</i>	Annahme: Die <i>Setzg</i> und die <i>SelbstZ</i> sind miteinander kompatibel. Aufgrund des kommissiven Charakters der <i>SelbstZ</i> gewinnt aber die <i>SelbstZ</i> .
6.1	<i>Dir</i>	<i>Setzg</i>	<i>Dir</i>	Annahme: Die <i>Dir</i> und die <i>Setzg</i> sind miteinander kompatibel. Aufgrund des kommissiven Charakters der <i>Dir</i> gewinnt aber die <i>Dir</i> .
6.2	<i>Setzg</i>	<i>Dir</i>	<i>Dir</i>	Annahme: Die <i>Setzg</i> und die <i>Dir</i> sind miteinander kompatibel. Aufgrund des kommissiven Charakters der <i>Dir</i> gewinnt aber die <i>Dir</i> .

⁹ „Position 1“ und „Position 2“ heißt: „Position 2“ ist – von der Textstelle mit der Verwendungsinstanz im Textverlauf zurückblickend – näher an der Verwendungsinstanz als „Position 1“.

7	<i>SelbstZ₁</i>	<i>SelbstZ₂</i>	<i>SelbstZ₂</i>	Annahme: Die beiden Definitionen sind miteinander kompatibel. Möglicherweise handelt es sich bei der zweiten aber um eine Präzisierung der ersten – daher gewinnt die zweite.
8	<i>Dir₁</i>	<i>Dir₂</i>	<i>Dir₂</i>	Annahme: Die beiden Definitionen sind miteinander kompatibel. Möglicherweise handelt es sich bei der zweiten aber um eine Präzisierung der ersten – daher gewinnt die zweite.
9	<i>Setzg₁</i>	<i>Setzg₂</i>	<i>Setzg₂</i>	Annahme: Die beiden Definitionen sind miteinander kompatibel (weil mit beiden ein universaler Gültigkeitsanspruch verbunden ist). Möglicherweise handelt es sich bei der zweiten aber um eine Präzisierung der ersten – daher gewinnt die zweite.
(10)	<i>FremdZ₁</i>	<i>FremdZ₂</i>	??	Für diesen Fall lässt sich keine Regel formulieren, die ausschließlich die Typenzugehörigkeit und die Auftretensreihenfolge berücksichtigt. Vielmehr ist in Fällen wie diesen zu ermitteln, ob sich irgendwo anders im Text (möglicherweise auch erst <i>nach</i> der Verwendungsinstantz) eine <i>SelbstZ</i> , eine <i>Dir</i> oder eine <i>Setzg</i> findet. Ist dies der Fall, dann ist diese als „Gewinner“ zu wählen (siehe auch nachfolgend Sonderregel 1). Ist dies nicht der Fall, d.h. enthält der gesamte Text nur <i>FremdZ</i> , dann kann nicht automatisch, sondern nur durch intellektuelle Textanalyse entschieden werden, welche Definition gewinnt.

Für Fälle, in welchen der Verwendungsinstanz eines Terminus überhaupt keine Definition oder ausschließlich *FremdZ* vorausgehen, wurde eine Sonderregel eingeführt, die den der Verwendungsinstanz *nachfolgenden* Text als Lieferant für mögliche infragekommene Definitionen zulässt. Für Fälle, in welchen mehr als zwei Definitionen miteinander konkurrieren, wurde eine Regelerweiterung formuliert, die beschreibt, nach welchem Procedere die Regeln jeweils paarweise auf eine beliebige Anzahl an Konkurrenten angewandt werden können:

Sonderregel:

Wenn einer Verwendungsinstanz eines Terms *X* entweder (a) keine Definition von *X* vorausgeht oder (b) lediglich ein oder mehrere Definitionen des Typs *FremdZ* vorausgehen, dann ist der Text *nach* der Verwendungsinstanz auf infrage kommende Definitionen abzusuchen. Die o.a. Regeln gelten dann (was die Reihenfolge des Auftretens der konkurrierenden Definitionen im Textverlauf betrifft) jeweils umgekehrt.

Erweiterung der Regeln auf beliebig viele konkurrierende Definitionen:

Wenn einer Verwendungsinstanz von *X* mehr als zwei Definitionen vorausgehen, wird wie folgt verfahren: Zunächst werden die ersten beiden gefundenen Definitionen in rückblickender Textverlaufsrichtung miteinander verglichen. Der „Gewinner“ wird mit der zunächst auftretenden Definition verglichen usw.

Beispiel:

Pos₁: Def₁(X) , Pos₂: Def₂(X) , Pos₃: Def₃(X) , Pos₄: Def₄(X) , Pos₅: TVI(X)¹⁰

Vorgehen:

- (1) Def₁(X) und Def₂(X) werden auf Basis der bekannten Regeln verglichen.
- (2) Der Gewinner aus (1) wird auf Basis der bekannten Regeln mit Def₃(X) verglichen.
- (3) Der Gewinner aus (2) wird auf Basis der bekannten Regeln mit Def₄(X) verglichen.
- ? Der Gewinner aus (3) ist die für TVI(X) verbindliche Definition.

Die Regeln wurden in einem in XSLT programmierten Ranking-Tool (*Def+Term2Rank.xsl*) implementiert¹¹, das zu jeder Termverwendungsinstanz eines gemäß der Schritte 1, 2 und 3 aufbereiteten XML-Dokuments eine hierarchisch geordnete Liste von Definitionen generiert, die dann als Basis für die automatische Generierung von Linkangeboten dienen kann, die in der Hypertext-Sicht auf ein Korpusdokument eine Termverwendungsinstanz mit dem jeweils für diese Verwendungsinstanz zuoberst gerankten definitorischen Textsegment verknüpfen.

3 Literatur

- Beißwenger, Michael, Lenz, Eva Anna & Storrer, Angelika (2002): Generierung von Linkangeboten zur Rekonstruktion terminologiebedingter Wissensvoraussetzungen. In: Stephan Busemann (Hrsg.): KONVENS 2002. 6. Konferenz zur Verarbeitung natürlicher Sprache. Proceedings, Saarbrücken, 30.09.-02.10.2002. Saarbrücken 2002 (DFKI Document D-02-01), 187-191.
- Beißwenger, Michael, Storrer, Angelika & Runte, Maren (2004): Modellierung eines Terminologienetzes für das automatische Linking auf der Grundlage von WordNet. In: Kunze, Claudia; Lemnitzer, Lothar; Wagner, Andreas (Hrsg.): Anwendungen des deutschen Wortnetzes in Theorie und Praxis. Beiträge des GermaNet-Workshops Tübingen, Oktober 2003 (LDV-Forum – Zeitschrift für Computerlinguistik und Sprachtechnologie 19. 1/2), 113-125.
- Budin, Gerhard (2000): Wissen(schaft)stheorie und Wissensorganisation. In: Ohly, H. Peter; Rahmsdorf, Gerhard; Sigel, Alexander (Hrsg.): Globalisierung und Wissensorganisation: Neue Aspekte für Wissen, Wissenschaft und Informationssysteme. Würzburg (Fortschritte in der Wissensorganisation 6), 41-48. Auch als Online-Veröffentlichung unter www.bonn.iz-soz.de/wiss-org/beitraege/Budin.doc.

10 Legende: *Pos* = Position (im linearen Textverlauf), *TVI* = Termverwendungsinstanz.

11 Programmierer: Benjamin Birkenhake.

- Enzyklopädie Philosophie und Wissenschaftstheorie. Hrsg. v. Jürgen Mittelstraß. Mannheim 1995f.
- Fluck, Hans-Rüdiger (1996): Fachsprachen. Einführung und Bibliographie. 5., überarb. u. erw. Aufl. Tübingen. Basel.
- Zifonun, Gisela, Hoffmann, Ludger & Strecker, Bruno (1997): *Grammatik der deutschen Sprache*. 3 Bde. Berlin. New York (Schriften des Instituts für deutsche Sprache 7.1-7.3).
- Gorski, D.P. (1967): Über die Arten der Definition und ihre Bedeutung in der Wissenschaft. In: Günter Kröber (Hrsg.): Studien zur Logik der wissenschaftlichen Erkenntnis. Übersetzung der russ. Originalausgabe von 1964. Berlin, 361-433.
- Kastberg, Peter (1999): Die Vertextung von Termini als Ausweis textueller Strategien in technischen Textsorten. In: *Hermes – Journal of Linguistics* 23, 41-63.
- Klavans, Judith & Muresan, Smaranda (2001): Evaluation of DEFINDER: A System to Mine Definitions from Consumer-Oriented Medical Text. In: Proceedings of the 1st JCDL. 2001. Roanoke, USA.
- Muresan, Smaranda & Klavans, Judith (2002): A Method for Automatically Building and Evaluating Dictionary Resources. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2002).
- Saggion, Horacio (2004). Identifying Definitions in Text Collections for Question Answering. In: Lino, Maria Teresa et al. (Hrsg.): Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisboa, Portugal 2004. [CD-ROM].
- Savigny, Eike von (1970): Grundkurs im wissenschaftlichen Definieren. München.
- Wiegand, Herbert Ernst (1996): Über usuelle und nichtusuelle Benennungskontexte in Alltag und Wissenschaft. In: Clemens Knobloch & Burkhard Schaefer (Hrsg.): Nomination – fachsprachlich und gemeinsprachlich. Opladen 1996, 55-103.