Lizenziatsarbeit der Philosophischen Fakultät der Universität Zürich

Institut für Computerlinguistik

Terminologische Definitionen: Form, Funktion, Extraktion

Referent:

Prof. Dr. Michael Hess

Verfasserin:

Mirjam Oberholzer

Mai 2002

Inhaltsverzeichnis

1 EINFÜHRUNG

1.1 Ziel

In der heutigen Informationsgesellschaft kommt der Handhabung von Wissen grosse Bedeutung zu. Ein Gebiet, das sich mit der Sammlung und Darstellung von Wissen beschäftigt, ist die Terminologie. Ihr Ziel ist die Gewinnung von Informationen zu Begriffen und deren einheitliche Repräsentation. Terminologie gelangt in verschiedenen Gebieten zum Einsatz. In Unternehmen fördert die Erstellung und Pflege einer Terminologiedatenbank den einheitlichen Sprachgebrauch. Die Sammlung von Informationen in einer zentralen Datenbank unterstützt zudem den Wissenstransfer innerhalb einer Firma. In Normungsgremien werden Begriffe geprägt und ihr Inhalt festgehalten. Dies erleichtert die Verständigung unter Experten eines Fachgebiets. Die oft mehrsprachige Terminologiearbeit internationaler Organisationen fördert einheitliche und unmissverständliche Übersetzungen.

Die Definition bildet innerhalb der Terminologiearbeit die zentrale Wissenseinheit. Sie hält den Inhalt des beschriebenen Begriffs fest und trägt damit zum Verständnis des jeweiligen Sachgebiets bei. Ziel der vorliegenden Arbeit ist es, die terminologische Definition unter verschiedenen Aspekten zu betrachten. Zum einen gelangt die Sicht der Terminologielehre zur Anwendung, die zulässige definitorische Formen diskutiert und eine Reihe von Anforderungen an Definitionen formuliert. Weiter soll die Umsetzbarkeit der theoretischen Kriterien am Beispiel einer vorhandenen Terminologiedatenbank geprüft werden. In diesem Zusammenhang finden sich auch Überlegungen zum Stellenwert der Definition in der übersetzungsorientierten terminologischen Tätigkeit. Definitionen sind gerade bei der in Verbindung mit Übersetzungstätigkeiten geleisteten Terminologiearbeit nicht immer greifbar. Aus diesem Grund beschäftigt sich die vorliegende Arbeit des Weiteren mit der automatischen Extraktion definitorischer Sätze. Dies ist ein Bereich, der nicht nur für die klassische Terminologiearbeit relevant ist, sondern auch im Gebiet der automatischen Wissensakquisition Anwendung finden kann. Mit Hilfe von lexikalisch-syntaktischen Mustern wird das informationstragende Umfeld von Begriffen identifiziert. Die so gewonnenen Informationselemente können zum Aufbau eines Wissenssystems verwendet werden.

Die vorliegende Arbeit gliedert sich in drei Teile. Der erste Teil (Kapitel 2) beschäftigt sich mit der Definition aus der Sicht der Terminologielehre. Die von der Lehre akzeptierten definitorischen Formen werden dargestellt und die inhaltlichen und formalen Anforderungen an eine richtig formulierte Definition erörtert. Der letzte Abschnitt des Kapitels 2 beschäftigt sich anhand einer konkreten Arbeitssituation mit der Relevanz der Definition für die praktische Terminologiearbeit. Diese Überlegungen werden im zweiten Teil (Kapitel 3) weitergeführt. Am Beispiel einer vorhandenen Terminologiedatenbank werden die Umsetzbarkeit der theoretischen Anforderungen und die Wichtigkeit der Definition geprüft. Der dritte Teil (Kapitel 4) beschäftigt sich mit der automatischen Identifikation definitorischer Sätze. Anhand verschiedener bereits implementierter Systeme wird die Extraktion von Definitionskandidaten englischer und französischer Sprache erörtert. Diese Diskussion bildet den Ausgangspunkt für die Entwicklung von Mustern zur Identifikation definitorischer Sätze im Deutschen. Anhand einiger implementierter Muster zeigen sich Anhaltspunkte zur Weiterentwicklung eines Programms für die Extraktion deutscher Definitionskandidaten.

1.2 Begriffserläuterungen

Im Folgenden sollen einige grundsätzliche Begriffe, die in der vorliegenden Arbeit Verwendung finden, kurz erörtert werden. Dazu gehören die Ausdrücke *Terminologie* und *Terminographie* an sich, aber auch die innerhalb der Terminologielehre gebräuchlichsten Bezeichnungen wie *Begriff, Begriffssystem, Merkmal* und *Definition* sowie *Benennung* und *Term*. Nachfolgend wird als Erstes die konzeptuelle Seite beschrieben (Begriff, Begriffssystem, Merkmal, Definition), dann die sprachliche Umsetzung der Begriffe (Benennung) und der Ausdruck *Term*.

Terminologie wird gelegentlich als Synonym von 'Terminologielehre' verwendet. Daneben bezeichnet Terminologie den Wortschatz einer Fachsprache sowie die Methoden der Terminologiearbeit.[1] Unter Terminographie wird einerseits die theoretische Auseinandersetzung mit den Prinzipien, Methoden und Verfahren zur Darstellung, Aufbereitung und Einordnung terminologischer Daten verstanden. Andererseits bezeichnet der Begriff die

[1] Mayer 1998, S. 25

praktische Terminologiearbeit, die sich mit der Tätigkeit des Erstellens und Bearbeitens von Termen mit entsprechenden Arbeitshilfsmitteln beschäftigt.[2]

Die zentrale Einheit der Terminologielehre ist der Begriff. Wüster versteht ihn als "das Gemeinsame, das Menschen an einer Mehrheit von Gegenständen[3] feststellen und als Mittel des gedanklichen Ordnens ('Begreifens') verwenden. Der Begriff ist somit ein Denkelement."[4] Felber definiert *Begriff* ebenfalls als eine durch Abstraktion gewonnene Einheit, ein "Denkgebilde", und weist vor allem auf die enge Beziehung zwischen Begriff und Merkmal hin:

> "Der Begriff ist ein Denkgebilde, das einem Gegenstand zugeordnet ist und diesen im Denken vertritt. Er ist eine Denkeinheit, die zur Bildung von logischen Sätzen (Aussagen) und zur Ordnung von Gegenständen dient. […] Der Begriff vertritt im strengen Sinn eine gleiche Menge von Eigenschaften, die einer Menge von Gegenständen eigen ist."[5]

DIN 2342 schliesst sich dem obigen Verständnis an und betrachtet den Begriff als eine "Denkeinheit, die aus einer Menge von Gegenständen unter Ermittlung der diesen Gegenständen gemeinsamen Eigenschaften mittels Abstraktion gebildet wird."[6] Zusammenfassend lässt sich sagen, dass der Begriff als Element verstanden wird, das durch eine Bündelung von Eigenschaften, den so genannten Merkmalen, gebildet wird. [7]

Merkmale können je nach Art der Betrachtung unterschiedlich klassifiziert werden. Anzutreffen sind vor allem Unterscheidungen nach inhärenten Merkmalen (*intrinsic characteristics*) und Relationsmerkmalen (*extrinsic characteristics*) oder nach wesentlichen und unwesentlichen Merkmalen.[8] Zur ersten Zweiteilung hält ISO 704 fest:

> "For the creation of new concepts in certain applied fields intrinsic characteristics such as shape, size, material, colour, position should be used in preference to extrinsic characteristics, such as origin, performance, location, discoverer, inventor."[9]

---

[2] Mayer 1998, S. 84
[3] DIN 2342 (1992, S. 1) definiert *Gegenstand* als "beliebigen Ausschnitt aus der wahrnehmbaren oder vorstellbaren Welt. Anmerkung: Auch Geschehnisse, Sachverhalte und Begriffe können Gegenstände sein".
[4] Wüster 1991, S. 8, zitiert nach: Mayer 1998, S. 30
[5] Felber 1990, S. 4
[6] DIN 2342 1992, S. 1
[7] Für Begriff wird im Englischen üblicherweise *concept* verwendet: "Concept: unit of thought constituted by those characteristics which are attributed to an object or to a class of objects". (ISO/DIS 1087 1988, S. 2, zitiert nach: Arntz et al. 1995, S. 43)
[8] Arntz et al. 1995, S. 56
[9] ISO 704 1987, S. 2, zitiert nach: Arntz et al. 1995, S. 56

Die zweite Klassifizierung betrachtet diejenigen Merkmale als wesentlich, die zur Definition oder Festlegung von Begriffsbeziehungen genützt werden. Alle übrigen fallen in die Kategorie der unwesentlichen Merkmale.[10] Diese Unterteilung kann allerdings nur im Kontext eines Begriffssystems (nachstehend erläutert) gemacht werden. Begriffssysteme lassen sich unter unterschiedlichen Gesichtspunkten zusammenstellen, und ein Merkmal kann je nach Organisation wesentlich oder unwesentlich sein.

Begriffsmerkmale und Beziehungen zwischen Begriffen bilden die Basis von Begriffssystemen. Die terminologische Theorie legt fest, dass Begriffe in einer systematischen Struktur präsentiert werden sollen.[11] Dazu werden sie einerseits aufgrund ihrer Merkmale, andererseits aufgrund der Beziehungen, die sie zu verwandten Begriffen haben, klassifiziert. Bei den Begriffsbeziehungen lassen sich hierarchische und nicht hierarchische unterscheiden. Zu den hierarchischen Beziehungen zählen die generische (auch *logische, Hyponymie-* oder *Abstraktionsbeziehung* genannt) und die partitive Beziehung (auch *Bestands-, Meronymie-* oder *Ganzes-Teil-Beziehung*).[12] Die generische Beziehung stellt eine hierarchische Ordnung her, indem sie Begriffe einem Oberbegriff zuordnet (z.B. *Personenkraftwagen* gehören zur Kategorie der *Kraftfahrzeuge*).[13] Die partitive Beziehung verdeutlicht die Verbindung zwischen einem mehrteiligen Begriff und seinen Konstituenten (z.B. ein *Rad* besteht aus *Nabe, Speichen, Felge*).[14]
Nicht hierarchische Beziehungen spielen in der Terminologielehre eine untergeordnete Rolle. Nach Arntz et al. können sie in sequenzielle und pragmatische Beziehungen unterteilt werden.[15] Sequenzielle Beziehungen charakterisieren sich durch das Zusammenspiel von Abfolgen. Zu den sequenziellen Beziehungen zählen beispielsweise chronologische (zeitliches Vor- und Nacheinander), Herstellungs- (Material/Produkt) oder Kausalbeziehungen (Ursache/Wirkung). Unter pragmatischen Beziehungen werden spezifische, gebietsabhängige Beziehungstypen verstanden. Sie beruhen auf "thematischen Zusammenhängen zwischen Begriffen", können jedoch "weder der hierarchischen noch der sequentiellen Begriffsbeziehung zugeordnet werden."[16]

[10] Arntz et al. 1995, S. 57
[11] Sager 1990, S. 28
[12] Mayer 1998, S. 34
[13] Arntz et al. 1995, S. 82
[14] Arntz et al. 1995, S. 53
[15] Arntz et al. 1995, S. 80
[16] DIN 2342 1986, S. 4, zitiert nach: Arntz et al. 1995, S. 81

Merkmale und Begriffssysteme bilden die Grundlage für die Begriffsbestimmung. Nach DIN 2342 ist die Begriffsbestimmung die "Festlegung eines Begriffs aufgrund seiner Merkmale im Rahmen eines Begriffssystems".[17] Die "Begriffsbestimmung mit sprachlichen Mitteln"[18] wiederum ist die Definition. Eine detaillierte Erörterung der terminologischen Definition und ihrer Formen und Merkmale folgt in Kapitel 2.

Die oben erläuterten Ausdrücke *Begriff*, *Merkmal*, *Begriffssystem* und *Definition* beschäftigen sich mit der konzeptuellen Ausprägung terminologischer Elemente. Nachfolgend soll auf die sprachliche Realisierung von Begriffen als Benennungen eingegangen werden. Begriffe als Denkeinheiten werden sprachlich wiedergegeben durch Bezeichnungen. Als Bezeichnungen gelten Benennungen, Symbole, Nummern und Notationen.[19] Benennungen sind die "lautsprachlichen Zeichen, mit deren Hilfe die Fachbegriffe ausgedrückt werden" und bilden die am häufigsten verwendete Bezeichnungsform.[20] DIN 2342 kennt Einwort- und Mehrwortbenennungen.[21] Die Zuordnung einer Benennung zu einem Begriff sollte wenn möglich eineindeutig oder wenigstens eindeutig sein.[22] Bei Eineindeutigkeit wird ein Begriff durch genau eine Benennung repräsentiert, bei Eindeutigkeit repräsentiert eine Benennung genau einen Begriff.[23] Bezeichnet eine Benennung mehrere verschiedene Begriffe, handelt es sich um Polysemie oder Homonymie. Von Polysemie spricht man, wenn eine Benennung in mehreren unterschiedlichen Bedeutungen, deren Zusammenhang noch erkennbar ist, verwendet wird (z.B. *Fuss* im Sinn von 'unterster Teil des Beines bei Mensch und Wirbeltier' und als 'Ständer oder Träger von Möbeln oder Gefässen'). Im Fall von Homonymie gleichen sich Benennungen in ihrer äusseren Form; die Begriffe, denen sie zugeordnet sind, weisen jedoch in ihrer Bedeutung keinerlei Ähnlichkeit auf (z.B. *Ton* im Sinn von 'Klang' und im Sinn von 'Erde').[24] Wird ein Begriff durch mehrere Benennungen repräsentiert, handelt es sich um Synonymie. Sager weist darauf hin, dass sowohl die eindeutige als auch die eineindeutige Beziehung zwischen Begriff und Benennung in der Praxis selten ist:

[17] DIN 2342 1992, S. 2
[18] DIN 2342 1992, S. 2
[19] Mayer 1998, S. 36
[20] KÜWES 1990, Kap. 2, S. 5
[21] DIN 2342 1992, S. 2
[22] Arntz et al. 1995, S. 130
[23] Mayer 1998, S. 11
[24] Arntz et al. 1995, S. 134-135

> "The recognition that terms occur in various linguistic contexts and that they have variants which are frequently context-conditioned shatters the idealised view that there can or should be only one designation for a concept and vice versa."[25]

Beim Begriff *Term* lassen sich unterschiedliche Verwendungen feststellen. DIN 2342 definiert ihn als "zusammengehöriges Paar aus einem Begriff und seiner Benennung als Element einer Terminologie".[26] Wüster[27] versteht darunter das fachsprachliche Synonym zu 'Benennung'.[28] Eine weitere Auffassung wird von Drozd et al. vertreten, die Term als ein Synonym von 'Fachwort' betrachten. Als Fachwörter bezeichnen sie "Wörter, denen […] eine wissenschaftliche Definition zugrunde liegt und die als direktere Zeichen einer bestimmten aussersprachlichen Wirklichkeit […] unter Ausschluss von semantischen Modifikationen verwendet werden."[29] Zusammenfassend lässt sich festhalten, dass zwar in Bezug auf die Auffassung des Begriffs *Term* ein gemeinsamer Nenner auszumachen ist – er repräsentiert einen Begriff, ist definiert und bezieht sich auf ein Fachgebiet –, aber eine allgemein anerkannte Definition nicht existiert. In der vorliegenden Arbeit wird *Term* im Sinn von 'Fachwort' verwendet.

---

25 Sager 1990, S. 58
26 DIN 2342 1992, S. 3
27 Wüster 1991, S. 36, zitiert nach: Mayer 1998, S. 37
28 Arntz et al. weisen darauf hin, dass diese Verwendung auch im Englischen häufig ist: "Term: designation of a defined concept in a special language by a linguistic expression" (ISO/DIS 1087 1988, S. 7, zitiert nach: Arntz et al. 1995, S. 40).
29 Drozd et al. 1973, S. 44, zitiert nach: Mayer 1998, S. 37

2 FORMEN, MERKMALE UND STELLENWERT DER TERMINOLOGISCHEN DEFINITION

Kapitel 2 beschäftigt sich mit der terminologischen Begriffsbestimmung. Zu diesem Zweck werden die von der Terminologielehre diskutierten Formen in Abschnitt 2.1.1 und 2.1.2 erläutert. Teil 2.1.3 diskutiert die Anforderungen, die an Definitionen gestellt werden. Die Relevanz von Definitionen für verschiedene Arten terminologischer Arbeit wird in Abschnitt 2.2 erörtert.

Zur Funktion der terminologischen Definition hält DIN 2330 fest:

> "Definitionen dienen dazu, einen möglichst eindeutigen Zusammenhang zwischen Begriffen und Benennungen herzustellen. Sie grenzen einen Begriff ab, indem er zu andern (bekannten oder bereits definierten) in Beziehung gesetzt wird."[30]

Eine terminologische Definition soll also einen Begriff innerhalb des dazugehörigen Systems eindeutig identifizieren. Die Terminologielehre geht davon aus, dass ein Begriff innerhalb eines Fachgebiets nur eine Bedeutung aufweist, und er wird als kontextfrei betrachtet.[31]

Die formale Struktur einer Definition wird üblicherweise als Gleichung dargestellt, wie dies Dahlbergs Definition vorwegnimmt:

> "A definition is the equivalence between a definiendum ('what is to be defined?') and a definiens ('how is something to be defined?') for the purpose of delimiting the understanding of the definiendum in any communication case."[32]

Die Gleichung enthält auf der linken Seite den durch eine Benennung repräsentierten Begriff, das Definiendum, und auf der rechten Seite die Begriffsbeschreibung, das Definiens. Zwischen beiden steht das Verbindungselement, das sie gleichsetzt.[33]

Sager weist darauf hin, dass eine terminologische Definition nicht unabhängig ist, da sie im Kontext sämtlicher Angaben eines terminologischen Eintrags betracht wird.[34] Ein Begriff ist Teil eines Begriffssystems, und zahlreiche Definitionen verweisen auf andere Elemente desselben Systems. So können beispielsweise Ober- und Unterbegriffe angeführt werden. Weiter kann der Gebrauch aussagekräftiger Kontexte eine Definition unterstützen.

---

[30] DIN 2330 1979, S. 9, zitiert nach: Arntz et al. 1995, S. 62
[31] Sager 1990, S. 41
[32] Dahlberg 1981, S. 17
[33] Arntz et al. 1995, S. 62
[34] Sager 1990, S. 45

Schliesslich sollte ein terminologischer Eintrag über eine Fachgebietsangabe verfügen, die den Gültigkeitsbereich der Definition auf das jeweilige Gebiet beschränkt. De Bessé betrachtet die Fachgebietsangabe als unerlässliche Begleitreferenz zu Definitionen, da sie die Verbindung zur extra-linguistischen Wirklichkeit schafft.[35] Aus dieser Einbettung in einem terminologischen Eintrag erklären sich auch die bevorzugten Definitionsformen und die Anforderungen, die an Begriffsbestimmungen gestellt werden.

2.1 Formen und Merkmale

2.1.1 Inhalts- und Umfangsdefinition

Die Inhalts- und die Umfangsdefinition sind die für die Terminologie relevantesten Definitionsarten.[36] Die Inhaltsdefinition, englisch *analytical definition, intensional definition* oder *generic definition* genannt[37], nimmt die aristotelischen Definitionsprinzipien auf und gibt *genus proximum* und *differentiae specificae* eines Begriffs an.[38] In der Inhaltsdefinition werden also, "ausgehend von einem bekannten bzw. bereits definierten Oberbegriff, die einschränkenden Merkmale angegeben, die den zu definierenden Begriff kennzeichnen und ihn von verwandten Begriffen unterscheiden".[39] Als Beispiel einer Inhaltsdefinition führt DIN 2330 folgende Definition an:

(1) *Glühlampe: ein materieller lichtaussendender Gegenstand (Oberbegriff), bei dem feste Stoffe durch Stromwärme so hoch erhitzt werden, dass sie Licht aussenden.*[40]

Ziel ist es nicht, alle bekannten Merkmale aufzunehmen, sondern nur diejenigen, die für den jeweiligen Zweck wesentlich sind, wobei die Art der benötigten Merkmale von der Platzierung des Begriffs im System abhängt.[41] Weist der Oberbegriff mehrere Unterbegriffe auf, kann das Definiendum durch die sie unterscheidenden Merkmale charakterisiert werden. Ist das Definiendum der einzige Unterbegriff, muss eines seiner inhärenten Merkmale angegeben werden, und zwar dasjenige, das seine Positionierung als Unterbegriff rechtfertigt. Die Auswahl der Merkmale sollte sich zudem nach dem Zielpublikum richten. Damit der Inhalt eines Begriffs verständlich wird, können je nach Spezialisie-

---

[35] de Bessé 1997, S. 67
[36] Sager 1990; Arntz et al. 1995; Felber 1993; KÜWES 1990; Dahlberg 1981
[37] Dahlberg 1981, S. 17; de Bessé 1997, S. 69
[38] Arntz et al. 1995, S. 66
[39] Arntz et al. 1995, S. 64
[40] DIN 2330 1979, S. 9, zitiert nach: Arntz et al. 1995, S. 64
[41] Sager et al. 1995, S. 62

rungsgrad des Lesers eines terminologischen Eintrags zusätzliche Merkmalangaben nötig sein.[42]

Die Inhaltsdefinition vermittelt durch die Anführung der begriffsbestimmenden Merkmale den Begriffsinhalt. DIN 2330 setzt Begriffsinhalt mit 'Intension' gleich und definiert ihn als "die Gesamtheit der Merkmale, die eine gedankliche Zusammenfassung von individuellen Gegenständen und die gegenseitige Abgrenzung der Begriffe ermöglichen".[43] Zudem ermöglicht die Inhaltsdefinition durch die Angabe der Merkmale und des Oberbegriffs die Einordnung des Konzepts in das Begriffssystem und die Abgrenzung gegenüber anderen Begriffen. Sager hält fest, dass die Terminologielehre einzig die Inhaltsdefinition anerkennen kann, da nur sie einen Begriff vollständig und systematisch identifiziert. Allerdings weist er auch darauf hin, dass in der Praxis wenige Definitionen diesen hohen Anforderungen genügen können:

> "In fact […] very few definitions have ever followed this strict pattern. A more relevant theory of terminology will have to admit the full range of definitions currently being used both in lexicography and terminology."[44]

Anzufügen ist auch, dass bei der punktuellen Terminographiearbeit der Begriff selten im Rahmen eines Begriffssystems erscheint. Entsprechend lassen sich in solchen Fällen Hyperonym und unterscheidende Merkmale kaum festmachen.

Im Gegensatz zur Inhaltsdefinition gibt die Umfangsdefinition den Begriffsumfang bzw. sämtliche zu einem Begriff gehörenden Gegenstände an. Der Begriffsumfang nach DIN 2342 entspricht der "Gesamtheit der einem Begriff untergeordneten Begriffe, die auf derselben Stufe stehen".[45] Der Umfang eines Begriffs kann auf unterschiedliche Art und Weise angegeben werden:

- Alle Unterbegriffe auf der gleichen Unterteilungsstufe werden aufgezählt:

  (2) *Unter "Bezeichnung" werden hier verstanden: Benennungen, Ideogramme, Nummern und Notationen.*[46]

- Alle individuellen Gegenstände werden genannt:

  (3) *Die Planeten des Sonnensystems sind Merkur, Venus, Erde, Mars, Jupiter, Saturn, Uranus, Neptun, Pluto*.[47]

---

[42] Rousseau 1983, S. 41
[43] DIN 2330 1979, S. 2, zitiert nach: Arntz et al. 1995, S. 48
[44] Sager 1990, S. 42
[45] DIN 2342 1986, S. 3, zitiert nach: Arntz et al. 1995, S. 50
[46] DIN 2330 1979, S. 8, zitiert nach: Arntz et al. 1995, S. 66
[47] DIN 2330 1979, S. 8, zitiert nach: Arntz et al. 1995, S. 66

Definitionen dieses Typs halten die Extension eines Begriffs fest. Im ersten Fall werden die Hyponyme, im zweiten die einzelnen Objekte aufgeführt. Die Umfangsdefinition ist weniger abstrakt als die Inhaltsdefinition und daher möglicherweise leichter verständlich. Sie kann aber nur dann ihren Zweck erfüllen, wenn die Anzahl der zum Begriff gehörenden Gegenstände begrenzt ist. Besonders im zweiten Fall verliert sie durch Veränderungen des Begriffs leicht ihre Gültigkeit.[48] Arntz et al. führen eine dritte Art der Umfangsdefinition an, die der Inhaltsdefinition ähnlich ist:

- Es wird die Regel angegeben, durch welche die Aufzählung gewonnen werden kann:

  (4) *Primzahlen werden daran erkannt, dass sie ausser durch eins und sich selbst durch keine andere ganze Zahl teilbar sind.*[49]

Dieser Typ könnte leicht in eine Inhaltsdefinition umgeformt werden, z.B. "Primzahlen sind Zahlen, die ausser durch eins und sich selbst durch keine andere ganze Zahl teilbar sind."

Zusammenfassend lässt sich sagen, dass die Inhalts- und Umfangsdefinition eng mit dem Begriffssystem verknüpft sind. Sie widerspiegeln dessen Ordnung, indem sie die für den Begriff relevanten hierarchischen Beziehungen verdeutlichen. Im Fall der Inhaltsdefinition ist dies die generische, bei der Umfangsdefinition die partitive Beziehung. Des Weitern werden für die Inhaltsdefinition Hyperonym und Merkmale und für eine Form der Umfangsdefinition die Hyponyme aufgeführt. Diese beiden Definitionsarten stellen also die terminologische Ordnung explizit dar.

2.1.2 Weitere Definitionsformen

Sager hält fest, dass sich die praxisorientierte Terminologielehre mit allen in der Terminologie und Lexikographie gebräuchlichen Definitionsformen auseinandersetzen muss, auch wenn diese den strikten Vorgaben für eine Inhaltsdefinition nicht genügen.[50] Die Klassifizierung weiterer Definitionstypen variiert je nach dem Aspekt der Betrachtung. Oft liegen verschiedenen Einteilungen keine grundlegenden inhaltlichen Unterschiede zugrunde, sondern nur andere Betrachtungsweisen. Strehlow stellt praxisnahe Kategorien vor. Er berück-

---

48 Arntz et al. 1995, S. 66

49 Arntz et al. 1995, S. 67
Fraglich ist, ob diese Definitionsform in Anbetracht ihrer Ähnlichkeit mit der Inhaltsdefinition tatsächlich der Umfangsdefinition zuzuordnen ist.

50 Sager 1990, S. 42

sichtigt auch *pragmatic definitions*, die den formalen und inhaltlichen Kriterien einer *well-formed definition* nicht genügen.[51] Dahlberg unterscheidet Definitionsarten nach der Art der Beziehung zwischen Begriffen, die sie verdeutlichen.[52] Felber[53] und KÜWES[54] beschränken sich auf die Beschreibung der Inhalts- und Umfangsdefinition und summieren andere Formen unter dem Begriff *Begriffserklärung* bzw. *Begriffsumschreibung*. Arntz et al. basieren ihre Auswahl von Definitionsarten auf deren Relevanz für die praktische Terminologiearbeit.[55] Sager betrachtet in seiner Klassifikation auch lexikographische und enzyklopädische Definitionstypen und Mischformen verschiedener Arten.[56] Sie entspricht zu einem grossen Teil de Bessés Unterteilung.[57] Nachfolgend werden die Arten nach Arntz et al., Sager und de Bessé diskutiert. Wenn möglich werden Querverbindungen geschaffen bzw. verwandte Definitionsformen gruppiert.

2.1.2.1 Nominaldefinition und Paraphrasierung

Arntz et al., de Bessé und Sager führen den Typ der Nominaldefinition (*definition by synonyms, définition par renvoi*) auf, die allerdings den terminologischen Anforderungen nicht entspricht.[58] Dabei wird ein unbekanntes Wort durch ein besser verständliches ersetzt (bspw. *bellis perennis* durch *daisy*). Die Nominaldefinition bezieht sich auf Wissen, das in der Sprache selbst verankert ist; es geht primär darum, die Benennung zu erklären.[59] Sie genügt den Anforderungen der Terminologielehre nicht. DIN 2330 hält fest, dass es nicht ausreicht, "lediglich eine Benennung durch eine andere zu ersetzen" und bezeichnet das als tautologische Definition.[60] De Bessé sieht in Synonymen die Möglichkeit, allgemeinsprachliche Wörter zu erklären, aber nicht Fachwörter zu definieren.[61] Echte Synonyme sind zudem in der Fachsprache selten und werden wenn möglich vermieden.[62] Arntz et al. attestieren der Nominaldefinition eine gewisse praktische Relevanz und räumen ein, dass sie trotz ihrer relativ geringen Aussagekraft dem Terminologen eine willkommene Hilfe sein kann.

---

[51] Strehlow 1983, S. 21
[52] Dahlberg 1981
[53] Felber 1993
[54] KÜWES 1990
[55] Arntz et al. 1995
[56] Sager 1990
[57] de Bessé 1990
[58] Arntz et al. 1995, S. 68; de Bessé 1990, S. 35; Sager 1990, S. 42
[59] Arntz et al. 1995, S. 68
[60] DIN 2330 1979, S. 8f., zitiert nach: Arntz et al. 1995, S. 68
[61] de Bessé 1997, S. 68
[62] Sager 1983, S. 122

Sager und de Bessé führen neben der Nominaldefinition die Definition durch Paraphrasierung (*definition by paraphrase, définition morpho-sémantique*) an, bei der beispielsweise *whiteness* als 'the state of being white' umschrieben wird.[63] Aufgrund ihres geringen Informationsgehalts ist sie für terminologische Zwecke nicht geeignet.[64]

2.1.2.2 Kontextdefinition

Die Kontextdefinition wird bei Arntz et al. und bei Sager beschrieben.[65] Sager bezeichnet sie als *definition by implication*. Bei der Kontextdefinition muss der Begriffsinhalt bzw. -umfang aus dem textlichen Zusammenhang erschlossen werden. Arntz et al. erachten diesen Definitionstyp als problematisch, da er auf ungenauen Angaben beruht, und führen als Beispiel die Kontextdefinition von *aircraft* in ISO/R 1087 an:

(5) *He went from Europe to America in 24 hours, using an […] aircraft.*[66]

Die beiden erschliessbaren Merkmale (Geschwindigkeit und Personenbeförderung) können den Kern des Begriffs Flugzeug nicht darstellen. Trotz ihrer Ungenauigkeit kann die Kontextdefinition dem Terminologen bei der Klärung von Begriffen aber eine nützliche Hilfe sein, da sich häufig keine Definitionen finden lassen, die den Ansprüchen an Begriffsbestimmungen im engeren Sinn genügen.[67]

2.1.2.3 Der Inhaltsdefinition verwandte Definitionsformen

Arntz et al. unterscheiden die genetische und die operationale Definition, die starke Ähnlichkeiten mit der Inhaltsdefinition aufweisen.[68] Die operationale Definition führt die einzelnen Operationen auf, die für die Bestimmung des Definiendums benötigt werden. So wird beispielsweise *Intelligenzquotient* durch die zu seiner Ermittlung dienenden Testoperationen definiert. Die genetische Definition definiert Vorgänge oder Ergebnisse von Vorgängen:

(6) *Protokoll: gleichzeitig erfolgende Niederschrift einer Verhandlung oder eines Verhörs.*[69]

---

[63] Sager 1990, S. 43; de Bessé 1990, S. 37
[64] de Bessé 1997, S. 69
[65] Arntz et al. 1995, S. 68; Sager 1990, S. 43
[66] ISO/R 1087 1969, S. 12, zitiert nach: Arntz et al. 1995, S. 68
[67] Arntz et al. 1995, S. 68
[68] Arntz et al. 1995, S. 67
[69] Arntz et al. 1995, S. 67

De Bessés *définition en compréhension* (auch *définition intentionnelle* oder *définition spécifique*) gibt die Merkmale eines Begriffs an und ähnelt in dieser Hinsicht der Inhaltsdefinition: "Cette définition donne l'ensemble des éléments sémantiques ou 'caractères' […] liés au défini." Als Beispiel figuriert die Definition von *lampe à incandescence* (Glühlampe):

(7) *Lampe électrique dans laquelle un matériau réfractaire est chauffé par un courant électrique à un point tel que la lampe commence à émettre de la lumière.*[70]

2.1.2.4 Beschreibende Definitionsformen

Sager führt eine Definitionsform an, die deskriptiver Art ist und daher enzyklopädische Züge aufweist. Er erläutert die *definition by synthesis* durch den Zusatz "by identifying relations, by description" und führt folgendes Beispiel an:

(8) *Metatarsalgia = a painful neuralgic condition of the foot, felt in the ball of the foot and often spreading thence up the leg*.[71]

Auch de Bessé kennt eine *définition dénotative* (oder *définition encyclopédique*), die beschreibend ist: "Cette définition est de type descriptif, substantiel. Elle décrit, dénote les éléments qui composent le défini."[72]

In die Kategorie der deskriptiven Definitionsformen lässt sich auch die in KÜWES aufgeführte *Begriffsumschreibung* einreihen. KÜWES hält fest, dass in manchen Sachgebieten Begriffsumschreibungen üblich sind, die den formalen und strukturellen Anforderungen an Definitionen nicht genügen. Dazu wird folgendes Beispiel angeführt:

(9) *Kaminfeger: Der Kaminfeger reinigt, wartet und kontrolliert wärmetechnische Anlagen und berücksichtigt dabei die Brandschutzvorschriften sowie die Erfordernisse der Lufthygiene, der Wärmewirtschaft und des Umweltschutzes.*[73]

2.1.2.5 Ostensive Definition

Sager weist auf die Möglichkeit der ostensiven Definition (*definition by demonstration*) hin.[74] Sie kann beispielsweise aus Zeichnungen und Photographien bestehen. Nach de Bessé kann diese Form für terminologische Zwecke geeignet sein: "The terminographer can also use illustration in order to define."[75]

---

[70] de Bessé 1990, S. 37
[71] Sager 1990, S. 43
[72] de Bessé 1990, S. 37
[73] KÜWES 1990, Kap. 4, S. 7
[74] Sager 1990, S. 43
[75] de Bessé 1997, S. 70

2.1.3 Anforderungen an Definitionen

Im vorherigen Abschnitt wurden verschiedene Definitionsarten diskutiert. Neben der Art einer Definition an sich lässt sich eine Reihe von inhaltlichen und formalen Anforderungen ausmachen, denen eine terminologische Begriffsbestimmung genügen soll. Sie sind nachfolgend dargestellt. Als Erstes werden die inhaltlichen Bedingungen erörtert, anschliessend die formalen Kriterien dargestellt.

2.1.3.1 Inhaltliche Anforderungen

Die folgenden Ansprüche werden an den Inhalt einer Definition gestellt:

- Die Definition soll sich an Zweck und Geltungsbereich orientieren[76] und dem Wissensstand und den Bedürfnissen des Zielpublikums entsprechen[77]. DIN 2330 illustriert diese Anforderungen mit folgendem Beispiel:

  > "Soll der Begriff *Zeitung* in einem allgemeinen Text definiert werden, um den Unterschied zum Fernsehen hervorzuheben, so reicht eine Definition etwa der Art: 'Zeitung ist ein auf Papier gedrucktes periodisch erscheinendes Massenmedium.' Für ein Lehrbuch des Bibliothekswesens, in dem *Zeitung* von Begriffen wie *Zeitschrift*, *Serienwerk* und *Buch* abgegrenzt werden muss, ist eine wesentlich genauere Definition notwendig (etwa: 'periodisch mindestens einmal wöchentlich erscheinendes Publikationsmittel mit universellem aktuellem Inhalt')."[78]

- Alle Begriffe eines Begriffssystems müssen unter dem gleichen sachlichen Gesichtspunkt definiert werden: "So wäre es falsch, im Begriffssystem "Huftiere" das Pferd als 'Reittier' (Gesichtspunkt: Verwendung) und das Rind als 'Paarhufer' (Gesichtspunkt: Zoologische Systematik) zu definieren."[79]
- Wenn möglich sollen im Definiens als Bezugspunkte verwendete Benennungen dem Begriffssystem des Definiendums entnommen werden. Dies dient der Verständlichkeit der Definition.[80]
- Definitionen müssen regelmässig aktualisiert werden. Da sich das menschliche Wissen in ständiger Entwicklung befindet, kann eine Definition nur so lange gültig sein, wie die Merkmale eines Begriffs unverändert bleiben.[81]

---

[76] Arntz et al. 1995, S. 70
[77] de Bessé 1997, S. 70
[78] DIN 2330 1979, S. 9, zitiert nach: Amtz et al. 1995, S. 71
[79] KÜWES 1990, Kap. 4, S. 6
[80] Arntz et al. 1995, S. 70
[81] Arntz et al. 1995, S. 71

- Zirkeldefinitionen sind zu vermeiden.[82] KÜWES ruft den Grundsatz in Erinnerung, dass ein Begriff nicht durch sich selbst definiert werden darf.

  (10) **Textilien = Produkte der Textilindustrie*
  Textilien = gewebte, gestrickte oder gewirkte aus Faserstoff hergestellte Waren[83]

  Als Spezialfall können allerdings Komposita betrachtet werden.[84] Der Kern eines Kompositums oder eines Mehrwort-Terms kann wiederholt werden, vor allem wenn es sich dabei um das Hyperonym handelt (z.B *an analog computer is a computer that […]*). Wird der Kern eines Kompositums als Oberbegriff verwendet, sollten jedoch die restlichen Wortelemente nicht in der Definition erscheinen.
- Negative Definitionen sind häufig fehlerhaft, weil das negative, ausschliessende Merkmal auch auf andere Begriffe zutrifft und ihm somit die einschränkende Wirkung abgeht.

  (11) **Anhänger = Fahrzeug ohne eigenen Antrieb*
  Anhänger = Fahrzeug, das für die Fortbewegung an ein Zugfahrzeug angehängt wird[85]

  Beim Definieren sollten nur dann negative Merkmale benutzt werden, wenn der Begriff selbst negativ ist.[86]
- Terminologische Definitionen sollten keine Redundanz aufweisen.[87] Sie sind in einen terminologischen Eintrag eingebettet und andere Datenkategorien wie Kontext oder Fachgebietsangabe liefern ergänzende Informationen.[88] Es ist auch unnötig, die inhärenten Merkmale des Hyperonyms erneut aufzulisten.

Bei den oben genannten Anforderungen lassen sich allgemein gültige und terminologiespezifische Bedingungen unterscheiden. Allgemein gültige Kriterien, die sich beispielsweise auch in der Lexikographie[89] finden, haben die Verständlichkeit und Eindeutigkeit der Definition zum Ziel, während terminologiespezifische Bedingungen die Konsistenz und Kohärenz des terminologischen Systems bezweckt. Orientierung an Zweck und Geltungsbereich, regelmässige Aktualisierung und das Vermeiden von zirkelhaften und negativen Definitionen sind Anforderungen, die auch an lexikographische Begriffserläute-

[82] Arntz et al. 1995, S. 72
[83] KÜWES 1990, Kap. 4, S. 7
[84] Ndi-Kimbi 1994, S. 333
[85] KÜWES 1990, Kap. 4, S. 7
[86] Arntz et al. 1995, S. 74
[87] Arntz et al. 1995, S. 75
[88] Sager 1990, S. 45
[89] Landau 2001, S. 157-171

rungen gestellt werden. Andere Bedingungen wie die Beibehaltung des sachlichen Gesichtspunkts, die Verwendung von systeminhärenten Benennungen und die Vermeidung von Redundanzen basieren auf der Einbettung der Definition in einem terminologischen Eintrag und in einem Begriffssystem. Diese Anforderungen können nur in der systematischen Terminologiearbeit erfüllt werden. Bei punktueller terminographischer Arbeit ist die Rahmenbedingung – ein ausgearbeitetes und kohärentes Begriffssystem – nicht erfüllt und entsprechend lassen sich die Bedingungen nicht umsetzen. Dieser Punkt wird in Teil 2.2.1 weiter ausgeführt.

2.1.3.2 Formale Anforderungen

Die formalen Ansprüche, die an terminologische Definitionen gestellt werden, haben primär ihre Kürze zum Zweck. Damit soll eine leichte Verständlichkeit sichergestellt werden.

- Das Definiendum in der Begriffsbestimmung aufzuführen ist nicht sinnvoll, da dies die Definition unnötig verlängert.[90]
- Die Definition sollte wenn möglich aus nur einem Satz bestehen.[91] Diese Angabe ist allerdings nur als Empfehlung zu verstehen, da es berechtigte Ausnahmen zu dieser Regel geben kann.

Der Abschnitt 2.1.3 behandelte die Anforderungen, die von der Terminologielehre an Definitionen gestellt werden. Unterscheiden lassen sich Kriterien, die bei sämtlichen Formen der Terminologiearbeit realisierbar sind und solche, die nur in der systematischen Tätigkeit umgesetzt werden können. Bei der Betrachtung des Stellenwerts von Definitionen wird häufig ebenfalls nach Tätigkeiten differenziert. Dieses Thema wird in Teil 2.2 erörtert.

2.2 Stellenwert der Definition in der terminologischen Arbeit

2.2.1 Stellenwert in der normenden, deskriptiven und übersetzungsorientierten Terminologiearbeit

Bei der Beurteilung der Wichtigkeit von terminologischen Definitionen muss zwischen deskriptiven und präskriptiven Tätigkeiten unterschieden werden. Deskriptive Termi-

[90] de Bessé 1997, S. 70
[91] de Bessé 1997, S. 70

nographiearbeit hat die Beschreibung und Aufzeichnung des Ist-Zustandes zum Ziel. Präskriptive oder normende Terminologiearbeit bezweckt die einheitliche Verwendung sowie die Eindeutigkeit bzw. die Eineindeutigkeit eines Fachwortschatzes.[92] Bei der deskriptiven oder beschreibenden Arbeit lassen sich punktuelle und systematische Terminologiearbeiten ausmachen. Punktuelle Terminologiearbeit bedeutet, dass ein einzelner Term untersucht wird. Falls dabei auch benachbarte Terme berührt werden, wird bestenfalls ein kleineres Begriffssystem erarbeitet. Bei der systematischen Terminologiearbeit wird ein gesamtes Fachgebiet bearbeitet. Dabei geht es einerseits um sachgebietsbezogene Untersuchungen, im Rahmen derer ein präzise eingeschränktes Gebiet terminologisch aufbereitet wird, andererseits um textbezogene Prüfungen, bei denen der in einem Text enthaltene Fachwortschatz bearbeitet wird.[93]

Als ein Spezialfall deskriptiver Arbeit kann die übersetzungsorientierte Terminologiearbeit betrachtet werden. DIN 2342 unterscheidet zwei Formen von terminologischen Tätigkeiten: die übersetzungsorientierte Arbeit, die in der Regel mehrere Sprachen behandelt, und die Einzelnormung, die sich auf eine Sprache bezieht.[94] Hohnhold versteht unter der übersetzungsorientierten Tätigkeit "mehrsprachige Terminologiearbeit, die Übersetzer befähigt, Fachübersetzungen als Fachsprache im Zusammenhang in der Zielsprache herzustellen."[95]

Die Grenzen zwischen präskriptiver Arbeit einerseits und deskriptiven und übersetzungsorientierten Tätigkeiten andererseits sind fliessend. Der normenden Arbeit geht üblicherweise ein beschreibendes Erfassen der Daten voraus. Weiter hat die präskriptive Terminologiearbeit zum Zweck, Benennungen und Begriffe festzulegen, und diese Tätigkeit beschränkt sich nicht nur auf die Arbeit von Normungsgremien:

> "Unter den Begriff *normende Terminologiearbeit* lässt sich […] auch die Terminologiearbeit fassen, die beispielsweise in zahlreichen grösseren Industrieunternehmen geleistet wird. Die so entstehende firmeneigene Terminologie spielt in der Praxis eine wichtige Rolle. […] Aber auch die Arbeit des Terminologen im Übersetzungsdienst hat häufig einen gewissen normenden Charakter, beispielsweise dann, wenn in einem behördlichen Sprachendienst die fremdsprachigen Äquivalente für die Bezeichnungen von Ämtern, Ministerien usw. verbindlich festgelegt werden."[96]

---

[92] Mayer 1998, S. 11
[93] Mayer 1998, S. 12
[94] DIN 2342 1992, S. 4
[95] Hohnhold 1982, S. 5
[96] Arntz et al. 1995, S. 233-234

Die Definition wird je nach terminologischer Tätigkeit unterschiedlich gewichtet. Bei der normenden Arbeit, im Rahmen derer Begriffe festgelegt werden, ist die Definition ein zentrales Element. Ohne sprachliche Festlegung seines Inhalts lässt sich ein Begriff nicht eindeutig fixieren. Bei der beschreibenden Arbeit hingegen sind die Begriffe bereits festgelegt und abgegrenzt. In diesem Fall ist die Definition primär für das Verständnis des Begriffs wichtig. Sie hilft auch bei der Klärung von Synonymie oder der Identifizierung von Lücken im Fachwortschatz. Die Definition ist zudem wesentlich, wenn durch eine Änderung des Begriffsinhalts eine Revision der Benennung notwendig wird.[97] Bei einigen Begriffen ist eine Definition jedoch nicht zwingend. Zahlreiche wissenschaftliche und technische Begriffe sind in der Fachliteratur bereits festgelegt und definiert. In diesen Fällen ist vorzugsweise die offizielle Formulierung zu übernehmen oder auf eine Definition ganz zu verzichten, wenn der Begriff als bekannt vorausgesetzt werden kann.[98] Andere Terme werden durch die Beziehungen, in denen sie zu anderen Begriffen stehen, identifiziert. Dadurch wird eine explizite Definition hinfällig. In diesem Zusammenhang betont Sager, dass die Definition nicht das einzige Mittel zur Begriffserläuterung ist.[99] Das Begriffssystem als solches hilft grundsätzlich bei der Identifizierung von Begriffen. Er weist auch darauf hin, dass der statische Ansatz von Definitionen die Dynamik der Wissensstruktur nicht reflektieren kann.[100] Vor allem die Inhaltsdefinition, die unterscheidende Merkmale anführt, müsste bei Systemänderungen konsequent angepasst werden. Auch Drozd weist darauf hin, dass Benennungen und Definitionen im Gegensatz zu Begriffsinhalten einen konservativ-statischen Charakter aufweisen.[101]

Bei der deskriptiven terminologischen Arbeit ist die Definition folglich unzweifelhaft von Nutzen, aber nicht immer notwendig. Innerhalb der beschreibenden Arbeit bildet die übersetzungsorientierte Terminologiebearbeitung ein Spezialgebiet, dessen Schwerpunkt bei der Identifizierung zielsprachlicher Äquivalente liegt. Auch in diesem Rahmen wird die Definition als wünschenswerte, aber nicht zwingende Information betrachtet. Dem Kontext wird aus praktischen Gründen die grössere Bedeutung beigemessen, wie Wright anmerkt:

> "Die Rahmenbedingungen der übersetzungsorientierten Terminologiearbeit erzwingen, dass Kontexte, die in der Quellsprache mit Sicherheit

---

97 Natanson 1983, S. 60 ff.
98 Sager 1990, S. 50
99 Sager 1990, S. 51
100 Sager 1990, S. 54
101 Drozd 1983, S. 92

> – öfter aber auch bei Paralleltexten in der Zielsprache – reichlich vorhanden sind, im Vordergrund stehen, während Definitionen verhältnismässig selten vorkommen."[102]

Hohnhold betrachtet die Definition als fakultative Datenkategorie bei der übersetzungsorientierten Arbeit.[103] Als grundlegende Informationseinheiten bei der Aneignung von Übersetzungskompetenz erachtet er neben der Benennung und der fachsprachlichen Wendung den Kontext. Er hält fest, dass der Kontext "den für die notwendige sprachganzheitliche Betrachtung wesentlichen phraseologischen Aspekt" einbringt und daneben die begriffliche Klärung erleichtern kann.[104] Hohnhold stuft also die genaue Darstellung des Begriffsinhalts in einem übersetzungsorientierten terminologischen Eintrag als zweitrangig ein. Primär muss die Verwendung des Begriffs geklärt werden. Nach Sager sollte der Fachübersetzer keine Definition für den Begriff in der Ausgangssprache benötigen, da er von einem Kontext ausgehend arbeitet.[105] Auch eine Definition für ein zielsprachliches Äquivalent ist seiner Ansicht nach nur nötig, wenn der Übersetzer sich der Verlässlichkeit der ihm zur Verfügung stehenden Hilfsmittel nicht sicher sein kann. Harris weist darauf hin, dass sich eine Definition erübrigt, wenn ein Fachwort eine einzige zielsprachliche Entsprechung aufweist, vorausgesetzt, dass eine Sachgebietsangabe vorhanden ist.[106] KÜWES hält fest, dass schon ein Eintrag, der nur eine Benennung und ihre Entsprechungen in anderen Sprachen sowie Quellenangaben enthält, wertvolle Dienste leisten kann.[107] Allerdings wird auch angemerkt, dass der Benutzer im Allgemeinen auf zusätzliche Angaben angewiesen ist. Die ÖNORM 2710, die zum Ziel hat, Fachübersetzern ein Minimum an Kriterien zur Verfügung zu stellen, um die für "übersetzerische Arbeit notwendige Terminologie in der erforderlichen Zuverlässigkeit zu erarbeiten, zu pflegen und zwecks schnellen und bequemen Zugriffs bereitzuhalten", führt die Definition neben dem Kontext als mögliche Begriffsbeschreibung an.[108] Eine Begriffsbeschreibung sollte in mindestens einer Form vorhanden sein, damit der terminologische Eintrag als ausreichend zuverlässig betrachtet werden kann.

---

[102] Wright 2001, S. 7
[103] Hohnhold 1982, S. 7
[104] Hohnhold 1982, S. 2
[105] Sager 1990, S. 49
[106] Harris 1983, S. 146
[107] KÜWES 1990, Kap. 4, S. 1
[108] ÖNORM 2710 1993, S. 2, zitiert nach: Mayer 1998, S. 114

Die Literatur ist sich einig, dass die Definition als ergänzende, fakultative Information dient. Die übersetzungsgerichtete Arbeit muss sich primär am Kontext orientieren, da nur er immer vorhanden ist. Eine genaue Klärung des Begriffsinhalts kann sich erübrigen, wenn die Verwendung der Benennung eindeutig ist. Sagers Argument jedoch, dass der Fachübersetzer keine Definition benötigt, trifft nur zu, wenn der Übersetzer sich tatsächlich auf einige wenige Gebiete spezialisieren kann. Ist er mit Texten unterschiedlicher fachlicher Herkunft konfrontiert, kann er für das grundlegende Textverständnis auf zusätzliche Informationen angewiesen sein. Um den Stellenwert der terminologischen Definition anhand einer konkreten Arbeitssituation zu beurteilen, wurde eine Umfrage bei zwanzig Übersetzern des Sprachendiensts der Credit Suisse in Zürich durchgeführt. Nachstehend folgt eine Beschreibung der dort vorhandenen Terminologiedatenbank und der Resultate der Befragung.

2.2.2 Untersuchung der Definitionen in der Credit Suisse-Terminologiedatenbank

Wie Mayer erläutert, basiert die übersetzungsorientierte Terminologiearbeit auf einer besonderen Ausgangssituation: Der Übersetzer hat einen (Fach-)Text vor sich, der in eine andere Sprache zu übertragen ist und der Begriffe enthält, deren Begriffsbedeutung und/oder deren Äquivalente in der anderen Sprache nicht bekannt sind.[109] Um den Text übersetzen zu können, muss der ungefähre Begriffsinhalt klar sein und das zielsprachliche Äquivalent mit der grössten begrifflichen Entsprechung gefunden werden. Übersetzungsorientierte Terminologiearbeit hat die Aufgabe, die für ein erfolgreiches Übersetzen von Fachtexten erforderlichen terminologischen Recherchen durchzuführen und die Ergebnisse festzuhalten, um das erarbeitete Wissen bei weiteren Übersetzungen nutzen zu können. Dies entspricht den Zielen der Terminologiearbeit bei der Credit Suisse. Die Terminologieabteilung ist Teil des Sprachendiensts. Innerhalb des Sprachendiensts werden bankspezifische Dokumente wie interne Weisungen, Informationen zu Bankprodukten oder Mitarbeiter- und Kundenzeitschriften übersetzt. Weiter gelangen auch Textsorten zur Übersetzung, die nicht das Kerngeschäft der Bank betreffen. Dazu gehören beispielsweise Applikations-Handbücher oder juristische Dokumente. Speziell im Rahmen von Marketingtexten wird eine Reihe anderer Fachgebiete angeschnitten, z.B. verschiedene

[109] Mayer 1998, S. 41

Sportarten oder Kulturanlässe. Die Terminologieabteilung beschäftigt sich zum einen mit der Abklärung von Begriffen aus den laufenden Übersetzungen, also mit punktuellen Recherchen, deren Ergebnisse in der Terminologiedatenbank festgehalten werden. Zum andern überprüft und ergänzt sie die von Übersetzern neu erstellten Einträge. Als drittes Ziel ist die systematische Erarbeitung der Terminologie einzelner Gebiete gesetzt.
Die Terminologiedatenbank soll die Übersetzer des Credit Suisse-Sprachendiensts unterstützen. Die terminologischen Informationen müssen in einer für die Übersetzer sinnvollen Art zur Verfügung gestellt werden. Besonders wichtig sind daher Benennungen in verschiedenen Sprachen und Erläuterungen in Form von Kontexten oder, falls greifbar, Definitionen. Die geprüften Einträge der Datenbank stehen zudem über eine Intranet-Schnittstelle sämtlichen Mitarbeitern der Credit Suisse zur Verfügung. In dieser Version soll sie bei der Übersetzung und dem Verständnis anderssprachiger Begriffe helfen, zum einheitlichen Sprachgebrauch innerhalb der Bank beitragen und Wissen zu bankspezifischen Begriffen vermitteln.

2.2.2.1 Quantität und Herkunft der Definitionen

Die Terminologiedatenbank weist insgesamt rund 29'000 geprüfte und fertig bearbeitete Einträge auf.[110] Davon enthalten 55% oder ca. 16'000 Einträge Definitionen in deutscher, englischer, französischer, italienischer oder spanischer Sprache. In diesen Einträgen finden sich rund 39'400 Definitionsfelder. Ein Eintrag mit Definitionsangabe weist folglich üblicherweise mehrere Definitionsfelder auf. Innerhalb eines Eintrags können Definitionen in verschiedenen Sprachen erscheinen. Auch innerhalb eines Sprachblocks sind Mehrfachdefinitionen möglich, obwohl das grundsätzlich als unnötig betrachtet wird. Eine Auswertung nach Sprachen zeigt, dass das Deutsche mit beinahe 15'500 von insgesamt 39'400 Definitionen die Mehrheit an Begriffsbestimmungen aufweist. Eine relativ grosse Häufigkeit ist auch bei englischen und französischen Definitionen zu finden (etwas über 9000 Vorkommen).

Zur Herkunft der gebräuchlichsten Definitionen lassen sich anhand der Quellenfelder Richtwerte ermitteln. Häufig verwendete Quellen werden in der Credit Suisse-Terminologiedatenbank mit einem 10-stelligen Code versehen, während vereinzelt gebrauchte Angaben als Volltext erfasst sind. Allerdings sind die Erfassung der Quellen und vor allem

[110] Stand per 31.12.2001

der Gebrauch der Codes nicht immer konsistent. Zudem können Quellenangaben fehlen. Aus diesen Gründen sind nur approximative Aussagen zur Herkunft der Quellen möglich.

Tabelle 1 Übersicht über die in der Credit Suisse-Terminologiedatenbank vorhandenen Definitionsfelder und die darauf bezogenen Quellenangaben

Feld	Eigenschaft	Häufigkeit	%
Definitionsfelder	insgesamt mit Quelle	39'404 35'670	100 90.5
Definitionsquellenfelder	insgesamt mit Code	35'670 31'026	100 87
Quellencode-Typen	insgesamt mit einer Häufigkeit über 100-mal	1195 47	100 4

Die Tabelle 1 zeigt eine Übersicht über die in der Credit Suisse-Terminologiedatenbank vorhandenen Definitionsfelder und die dazugehörenden Quellenangaben. Die Auswertung der Definitions-Quellenfelder zeigt, dass bei 35'670 der insgesamt 39'404 Definitionen eine Quelle angegeben ist. Bei 90.5% aller Definitionen ist also die Herkunft belegt. Bei den vorhandenen Angaben sind 31'026 Quellen (87%) codiert. Die 31'026 codierten Quellen verteilen sich auf 1195 unterschiedliche Codes. 7 dieser 1195 Quellencodes erscheinen über 500-mal, 40 weisen eine Häufigkeit zwischen 100 und 499 Vorkommen auf.[111] Insgesamt decken diese 47 Codes 21'079 Quellen ab.

Bei den 47 Quellen, die über 100-mal verwendet werden, lassen sich folgende Herkunftstexte ausmachen:

- externe Handbücher oder Instruktionstexte (z.B. ISO-Normen) (13 Quellen)
- interne Handbücher (11 Quellen)
- externe Wörterbücher (9 Quellen)
- interne Glossare (6 Quellen)
- andere interne Quellen (z.B. Geschäftsbericht, Weisungen) (4 Quellen)
- Gesetzestexte (2 Quellen)
- Übersetzer und Terminologen (2 Quellen)

Die übersetzungsorientierte Ausrichtung des Terminologiedienstes spiegelt sich in der Tatsache, dass nur 55% aller Einträge eine Definition aufweisen. Der Schwerpunkt liegt bei der punktuellen Recherchenarbeit, im Rahmen derer Definitionen nur falls mit wenig Aufwand auffindbar eingefügt werden. Definitionen werden in erster Linie bei der systema-

[111] Weitere 272 Quellen figurieren zwischen 10- und 99-mal. Die überwiegende Mehrheit (876 Quellen) erscheint zwischen ein- und neunmal.

tischen Terminologiearbeit innerhalb von Projekten erfasst. Das kann die relativ geringe Anzahl der verwendeten Definitionsdokumente (1195 unterschiedliche Texte für 31'026 Quellen) erklären. Zudem stammen rund 21'000 Definitionen aus nur 47 unterschiedlichen Quellen. Externe und interne Dokumente halten sich bei den Quelltexten die Waage. Die übersetzungsorientierte Arbeit findet nicht in einem geschlossenen Rahmen statt und stützt sich entsprechend oft auch auf externe Quellen. Bei der Terminologiearbeit, die sich mit intern verwendeten Benennungen beschäftigt und häufig auch die Prägung neuer fremdsprachiger Benennungen umfasst, können Kurzerläuterungen zur Klärung eingegeben werden (z.B. *Projekt der CSFS* oder *Bezeichnung einer Organisationseinheit auf Stufe Ressort*). Dies ist sicher mit ein Grund, dass Übersetzer und Terminologen zu den häufigsten Quellen gehören.

2.2.2.2 Stellenwert der Definition aus der Sicht von Fachübersetzern

Ziel der Umfrage war es, die Meinung von Fachübersetzern zur Relevanz der Felder eines Terminologieeintrags einzuholen. Insbesondere die Wichtigkeit von Definitionen (unter anderem auch im Vergleich zu Kontexten) sollte festgestellt werden. Neben Angaben zur Qualität der Credit Suisse-Terminologiedatenbank beantworteten zwanzig Übersetzer die nachstehenden Fragen wie folgt:

1. Welche Felder – neben den Benennungen – sind am relevantesten innerhalb eines Eintrags?

   Zur Auswahl standen Fachgebietsangabe, Quellenangabe, Definition, Kontext, Anmerkungen (inhaltlicher oder linguistischer Natur), Aktualitätsangabe. Eine Mehrfachauswahl war möglich.

   Definitionen (17 Stimmen) wurden am häufigsten genannt. Als relevant wurden auch Kontext und Quellenangabe (je 13-mal) und Fachgebiet (12) eingestuft. Zweimal wurde auch der Autor eines Eintrags (ein systemgeneriertes Feld) als wichtige Information angeführt, was die Wichtigkeit der Herkunftsangabe bzw. Quelle unterstreicht.
2. Welche Felder sind weniger wichtig?

   Zur Auswahl standen wie oben Fachgebietsangabe, Quellenangabe, Definition, Kontext, Anmerkungen, Aktualitätsangabe. Auch hier konnten mehrere Punkte gewählt werden.

Als weniger wichtig eingestuft wurden vor allem Anmerkungen (14). Einige Punkte erhielten auch die Aktualitätsangabe (7), das Fachgebiet (4) und der Kontext (4).

3. Welche Felder sollten mehr Information enthalten?
   Zur Auswahl standen Quellenangabe, Definition, Kontext und Anmerkungen (Mehrfachauswahl möglich).
   Zusätzliche Informationen wurden vor allem für das Definitions- (10) und das Kontextfeld (5) gewünscht. 3 Punkte erhielt auch die Quellenangabe.
4. Welche Information ist nützlicher: Kontext oder Definition?
   15 Übersetzer fanden eine Definition nützlicher, vier einen Kontext.
5. Wie hoch (auf einer Skala von 1 bis 10) ist die Nützlichkeit eines Eintrags, der weder Kontext noch Definition enthält?
   (1 = schlecht, 4 = zufriedenstellend, 7 = gut, 10 = ausgezeichnet)
   Der Durchschnittswert lag bei 3.3 Punkten.

Zwei weitere Fragen zur Qualität der Credit Suisse-Terminologiedatenbank lassen indirekt Rückschlüsse auf die Relevanz von Definitionen zu:

6. Wie ist die Beurteilung des Informationsgehalts der Einträge?
   Zur Auswahl standen sehr gut, gut, zufriedenstellend, mangelhaft.
   Die meisten Übersetzer stuften den Informationsgehalt als gut (10) bis zufriedenstellend ein (8).
7. Was wird an der Terminologiedatenbank am meisten geschätzt?
   Als mögliche Punkte waren Vielsprachigkeit, Informationsgehalt, Quantität der Einträge und Abfragekomfort aufgeführt.
   Der Informationsgehalt wurde am häufigsten genannt (11), gefolgt vom Abfragekomfort (8) und der Vielsprachigkeit sowie der Quantität der Einträge mit je 5 Stimmen.

Abweichend von der in Abschnitt 2.2.1 erörterten Meinung der Literatur wird die Definition von Fachübersetzern als relevante Datenkategorie eingestuft, der gegenüber dem Kontext den Vorrang gegeben wird. Die Definition erscheint am häufigsten bei der Frage nach wichtigen Feldern und figuriert nicht unter den als unwichtig eingestuften Informa-

tionen. Zusätzliche Angaben werden ebenfalls am häufigsten bei den Definitionsfeldern gewünscht. Eine Mehrheit der Übersetzer zieht die Definition im terminologischen Eintrag dem Kontext vor. Eine Übersetzerin merkte dazu an, dass ein Kontext nur in der Zielsprache, eine Definition aber in jeder Sprache nützlich ist. Zweifellos ist jedoch je nach Art des Begriffs eine Definition nicht nötig, was der Durchschnittswert von 3.3 Punkten für die Nützlichkeit eines Eintrags ohne Definition oder Kontext vermuten lässt. Ein Eintrag ohne diese Angaben wird nur knapp als ungenügend eingestuft. Einige Übersetzer wiesen darauf hin, dass übermässig ausgestaltete Einträge (bspw. mit mehreren Definitionen oder Kontexten innerhalb eines Sprachblocks) sehr zeitraubend sind. Angemerkt wurde auch, dass je nach Art des Begriffs eine Definition oder ein Kontext unnötig sind. Als Beispiel wurden Eigennamen angeführt (z.B. *Schweizerischer Verband der Immobilien-Treuhänder*). In diesen Fällen reicht das fremdsprachige Äquivalent ohne jegliche Erklärung für die Übersetzung aus. Allerdings wurde auch festgehalten, dass im Fall von Homographen eine Unterscheidung zwischen verschiedenen Begriffen ohne zusätzliche Angabe schwierig ist. Die Beurteilung des Informationsgehalts der Datenbank liegt zwischen genügend und gut. Zahlreiche Rückmeldungen wiesen darauf hin, dass einige der Einträge sehr gut ausgebaut sind, während andere nur rudimentäre Informationen enthalten. Das entspricht dem Bild, das die Auswertung der Definitionsfelder zeigte. Begriffsbestimmungen sind zwar zahlreich, konzentrieren sich jedoch auf nur 55% aller Einträge. Der Informationsgehalt wird grundsätzlich mehr geschätzt als die Vielsprachigkeit der Datenbank. In Verbindung mit den oben stehenden Aussagen lässt dies darauf schliessen, dass auch bei Fachübersetzern ein Bedürfnis nach Begriffserläuterungen besteht.

Kapitel 2 behandelte die Formen und Merkmale von Definitionen aus Sicht der Terminologielehre. Weiter wurde der Stellenwert der Definition in verschiedenen terminologischen Tätigkeiten aus der Sicht der Literatur und anhand einer Befragung von Fachübersetzern erörtert. Die Wichtigkeit von Definitionen und die Umsetzbarkeit der an sie gestellten Anforderungen sollen im folgenden Kapitel 3 anhand der Evaluation einer Terminologiedatenbank geprüft werden

3 UNTERSUCHUNG DER SIEMENS-TERMINOLOGIEDATENBANK: KONSISTENZ, KOHÄRENZ UND QUALITÄT DEFINITORISCHER ELEMENTE

Die Firma Siemens in Karlsruhe unterhält eine Terminologie-Datenbank (SIMATIC-TermDB) für die technische Redaktion und für Übersetzungen. Ziel des nachstehend beschriebenen Projekts war es, die Einträge der SIMATIC-TermDB mit einem genormten Terminologiebestand zu vergleichen. Besonderes Gewicht wurde den in der SIMATIC-TermDB vorhandenen Definitionen und definitorischen Elementen zugemessen, da es sich dabei um die für den inhaltlichen Abgleich wichtigsten Daten handelt.

Die nachfolgende Projektbeschreibung konzentriert sich einerseits auf die zur Evaluation einer bestehenden Terminologiedatenbank (d.h. der SIMATIC-TermDB) verwendete Systematik (Kapitel 3.1), andererseits auf die Quantität und Qualität definitorischer Elemente innerhalb dieser Datenbank (Kapitel 3.2). Die in Kapitel 2.1 diskutierte Literatur bildet die Basis der Gesamtbeurteilung. Gleichzeitig ermöglichte die Untersuchung der Datenbank Rückschlüsse auf die Umsetzbarkeit der theoretischen Anforderungen und die Nützlichkeit der Felder eines terminologischen Eintrags.

3.1 Systematik

Ein genormter Bestand bildete den Referenzpunkt des Projekts. Er wird in Abschnitt 3.1.1 beschrieben. Als erste vorbereitende Arbeit wurde eine Statistik der SIMATIC-TermDB erstellt. Sie bietet die Möglichkeit, eine strukturierte Übersicht über die umfangreiche Datenmenge zu gewinnen (Abschnitt 3.1.2). Zu Beginn des Vergleichs der SIMATIC-TermDB mit dem genormten Bestand zeigte sich zudem, dass die Benennungen des SIMATIC-Bestands einige Inkonsistenzen aufweisen. Mit probeweisen Abgleichen der Daten wurden sie identifiziert und der SIMATIC-Bestand wie in Abschnitt 3.1.3 beschrieben leicht modifiziert. Die Vorgehensweise beim eigentlichen Datenvergleich ist in Abschnitt 3.1.4 dargestellt.

3.1.1 Genormter Bestand als Referenzpunkt

Der einleitend erwähnte genormte Bestand ist Teil der von der International Electrotechnical Commission (IEC) veröffentlichten Terminologie auf CD-ROM. Die IEC-CD-ROM enthält Begriffe zu rund 70 Sachgebieten, unter anderem zu *Automatic Control*.

Dieser Teilbestand wurde für den Vergleich mit den SIMATIC-Daten verwendet. Das Gebiet *Automatic Control* behandelt 342 Begriffe. Die Benennungen sind in Französisch, Englisch, Russisch, Arabisch, Deutsch, Spanisch, Japanisch, Polnisch, Portugiesisch und Schwedisch vorhanden. In Englisch, Französisch und Russisch sind Definitionen zu den 342 Begriffen angegeben. Gelegentlich sind Synonyme und Abbildungen integriert.
Der SIMATIC-Bereich der Firma Siemens befasst sich mit Produkten und Systemen, die für die Lösung von Automatisierungsaufgaben benötigt werden. Die IEC ist eine globale Organisation, die internationale Normen für alle elektrischen, elektronischen und verwandten Technologien vorbereitet und publiziert. Der IEC-Bestand zu *Automatic Control* ist daher inhaltlich relevant für die SIMATIC-TermDB. Einerseits ermöglicht der Vergleich zwischen den beiden Datenbeständen eine Beurteilung der Konformität des SIMATIC-Bestandes mit internationalen Richtlinien, andererseits können fehlende Begriffe eingegliedert werden. Für das Terminologieprojekt selbst ist die Arbeit mit einem genormten Bestand in verschiedener Hinsicht vorteilhaft. Die IEC-Daten bieten als in sich geschlossenes Thema und durch die für jeden Eintrag vorhandene Definition die Möglichkeit, ein Basisverständnis des Sachgebiets aufzubauen. Da die Definitionen sorgfältig formuliert sind, können sie als Referenzpunkt für die inhaltliche Bewertung der SIMATIC-Erläuterungen dienen.

3.1.2 Statistik

Die SIMATIC-TermDB steht als Textdatei zur Verfügung, die im MultiTerm-Exportformat strukturiert ist. Das MultiTerm-Exportformat basiert auf einer in der Terminologiesoftware *MultiTerm* der Firma Trados enthaltenen Standard-Exportdefinition. Die Einträge werden dabei in der folgenden Form exportiert (Beispiel der SIMATIC-TermDB entnommen):

```
**
<Eintragsnummer>42334
<Fachgebiet>SIMATIC S7
<ITA>(ri)caricare
<Abgestimmt>ja
<Status>gesperrt
<Bemerkung>(ri)caricare war ursprünglich als Synonym eingetragen.
<GER>booten
<Abgestimmt>ja
<Status>bevorzugt
<Last updated by>TDB
<Last updated>01.08.2000
<Definition>Ladevorgang, der das Betriebssystem vom Systemdatenträger in den
Systemspeicher überträgt.
**
```

Der Eintrag ist an Anfang und Ende durch zwei Asteriske begrenzt. Alle Felder eines Eintrags werden exportiert. Der Feldname ist durch eckige Klammern gekennzeichnet und wird direkt vom Feldinhalt gefolgt. Diese strukturierte Darstellung ermöglichte eine einfache automatisierte Bearbeitung durch Skripte in der Programmiersprache Perl.
Erstes Ziel der Statistik war es, sämtliche in der Datenbank vorhandenen Feldtypen und die Häufigkeit ihres Vorkommens zu identifizieren. Bei Feldern mit vordefiniertem Inhalt (Auswahlfelder) wurden zudem die Feldwerte extrahiert. Daraus lässt sich ein prototypischer Minimaleintrag der Terminologiedatenbank ableiten, der die folgende Form aufweist:

- Eintragsnummer (systemgeneriert und eindeutig)
- Fachgebietsangabe (Auswahlfeld, in allen Einträgen vorhanden)
- Benennung (in italienischer, französischer, englischer, spanischer oder deutscher Sprache, Mehrfachvorkommen möglich)
- "Abgestimmt" (ein Auswahlfeld mit den Möglichkeiten *ja* oder *nein*, mit Bezug auf die Benennungsfelder)
- Statusangabe (ein Auswahlfeld mit den Möglichkeiten *bevorzugt*, *gesperrt* und *frei*, mit Bezug auf die Benennungsfelder)
- "Last updated by" (systemgeneriert, enthält den für die Aufdatierung verantwortlichen Benutzer)
- "Last updated" (systemgeneriert, enthält das Aufdatierungsdatum)

Neben den oben genannten Feldern, die in allen Einträgen vorhanden sind, existieren andere Angaben. Zusätzlich zur Auswahl des Fachgebiets ist die Bestimmung eines Teilgebiets möglich. Daneben sind die Felder Grammatik, Benennungsquelle, Definition und Definitionsquelle, Kontext mit zugehöriger Quelle sowie ein Bemerkungsfeld vorhanden. Es existieren einige wenige Benennungen in weiteren Sprachen (Norwegisch, Polnisch, Schwedisch) und ein für alle Sprachen gültiges Benennungsfeld "Abkürzung/Vollbegriff". Zu den Feldern mit vordefiniertem Inhalt zählen Sach- und Teilgebietsangabe, das Feld "Abgestimmt" und die Statusangabe.

3.1.3 Modifikation des SIMATIC-Bestandes

Eine genauere Betrachtung der Benennungen der zwei zu vergleichenden Datenbestände zeigte einige Unterschiede formaler Art, die Modifikationen nötig machten. Dazu gehört

zum einen das Zeichen Eszett (ß). Es wird in den beiden Beständen unterschiedlich verwendet und wurde aus diesem Grund durch Doppel-s ersetzt. Weiter liess eine Analyse der Feldwerte zu "Abkürzung/Vollbegriff" darauf schliessen, dass das Feld sprachunabhängig verwendet wird. Der Feldinhalt umfasst sowohl Abkürzungen als auch Vollbenennungen verschiedener Sprachen. Um einen nach Sprachen geordneten Vergleich der IEC- und SIMATIC-Benennungen zu ermöglichen, wurde die Feldbezeichnung "Abkürzung/Vollbegriff" in den entsprechenden Sprachfeld-Namen (Deutsch, Englisch, Französisch, Spanisch oder Italienisch) überführt.
Schliesslich erwiesen einige probeweise vorgenommene Abgleiche, dass ein Teil der SIMATIC-Benennungen Sonderzeichen enthält. Gefunden wurden runde Klammern, Komma, Zirkumflex und Strichpunkt. Benennungen mit Zirkumflex oder Strichpunkt waren selten und wurden daher nicht untersucht. Eine Überprüfung der runden Klammern zeigte, dass sie häufig zur Darstellung von Synonymen benützt werden. So stellt beispielsweise *GAP-(Aktualisierungs)faktor* die zwei Synonyme *GAP-Aktualisierungsfaktor* und *GAP-Faktor* dar. Weiter können Klammern Schreibvarianten ausdrücken (*Alle(s) Löschen*) oder Zusatzinformationen (*Ganzzahl (16 Bit)*) anzeigen. Synonyme werden teilweise aber auch als eigenständige Benennungen im Sprachfeld erfasst, wie das im IEC-Bestand der Fall ist. Das Komma trennt häufig das Substantiv vom nachgestellten Adjektiv (*Sternkoppler, aktiv*). In einigen Mehrwortbenennungen wird das Adjektiv jedoch vorgestellt (*manueller Neustart mit Gedächtnis*), was der IEC-Usanz entspricht. Das Komma kann auch zwei Synonyme trennen (*wichtige Meldung, Systemfehlermeldung*).

Für den Abgleich der deutschen Benennungen wurde der Inhalt des entsprechenden SIMATIC-Sprachfelds in zusätzlicher Form generiert mit dem Ziel, alle möglichen Übereinstimmungen zwischen den SIMATIC- und den IEC-Daten zu finden.

- Runde Klammer

  Das Ziel der zusätzlich generierten Benennungen war es einerseits, in Klammern geschriebene Synonymausdrücke aufzulösen. Andererseits sollten Zusatzinformationen aus dem Benennungsfeld gelöscht werden. Die Klammer wurde aus der Originalbenennung entfernt, was bei Synonymausdrücken die Benennung der maximalen Länge ergab (12a). Daneben wurde eine Benennung ohne Klammerteil generiert (12b). Damit sollten unnötige Benennungsteile oder Zusatzinformation gelöscht werden.

(12) Original: *GAP-(Aktualisierungs)faktor*
generiert: a) *GAP-Aktualisierungsfaktor*
b) *GAP-faktor*

- Komma

  Bei Kommata wurde die Behandlung von Substantiv-Adjektiv-Konstruktionen angestrebt. In den zusätzlich generierten Benennungen sollte das Adjektiv dekliniert und vor dem Substantiv platziert werden. Zu diesem Zweck wurden neben der Originalbenennung vier zusätzliche Benennungen geschrieben, in denen der zweite Teil vorangestellt war (13a). In drei Fällen wurde der zweite Teil mit verschiedenen Deklinations-Endungen ergänzt (*-e/-es/-er*) (13b bis 13d).

(13) Original: *Sternkoppler, aktiv*
generiert: a) *aktiv Sternkoppler*
b) *aktive Sternkoppler*
c) *aktives Sternkoppler*
d) *aktiver Sternkoppler*

3.1.4 Datenvergleich

Der Abgleich des genormten IEC-Bestandes von 342 Einträgen mit der SIMATIC-TermDB wurde in drei Schritten durchgeführt:

1. Es wurde eine Liste der sowohl in der SIMATIC-TermDB als auch im IEC-Bestand vorhandenen deutschen Benennungen generiert.
2. Bei den in beiden Beständen vorhandenen deutschen Benennungen wurde überprüft, welche Benennungen in den IEC-Daten fremdsprachige Entsprechungen (in Englisch, Französisch, Spanisch, Italienisch) aufweisen, die nicht denjenigen der SIMATIC-TermDB entsprechen.
3. Bei den in beiden Beständen vorhandenen deutschen Benennungen wurde durch eine manuelle Untersuchung abgeklärt, ob eine begriffliche Übereinstimmung zwischen den IEC- und SIMATIC-Daten vorhanden ist.

Bei Einträgen ohne Definitionen gestaltete sich die Beurteilung der inhaltlichen Übereinstimmung schwierig und konnte nur mit Vorbehalt gemacht werden. Das nachfolgend erläuterte Bewertungsschema wurde als Anhaltspunkt für die Bestimmung der Begriffsdeckung verwendet. Folgende Kriterien resultierten nach Häufigkeit ihres Zutreffens in der Beurteilung "gleicher Begriff" oder "vermutlich gleicher Begriff":

- Die IEC-Definition findet sich in Informatik-Wörterbüchern.[112] Der Begriff wird daher als verbreitet betrachtet und die Wahrscheinlichkeit der Übereinstimmung mit dem SIMATIC-Eintrag als höher eingestuft.
- In den einschlägigen Wörterbüchern entspricht die Benennung nur einem Begriff. Da dies die Wahrscheinlichkeit von Homographen verringert, unterstützt dieses Kriterium die Vermutung, dass es sich um zwei Begriffe mit demselben Inhalt handelt.
- Die Fachgebietsangabe eines SIMATIC-Eintrags oder eine Notiz im Bemerkungsfeld weisen auf ein mit dem IEC-Eintrag verwandtes Sachgebiet hin. Die Wahrscheinlichkeit der begrifflichen Übereinstimmung bei gleichen Benennungen ist innerhalb eines Gebiets grösser. In einigen wenigen Fällen wird explizit auf IEC-Dokumente verwiesen, was als Beleg für eine inhaltliche Deckung gewertet wurde.
- In den fremdsprachlichen Benennungen bestehen in mindestens zwei Sprachen Übereinstimmungen. Auch dies wurde als Indiz für eine Begriffsübereinstimmung gewertet.

3.2 Resultate

Das folgende Kapitel behandelt die Ergebnisse des Projekts. In Abschnitt 3.2.1 werden die Anhaltspunkte diskutiert, die die vorbereitenden Arbeiten zur Beurteilung von Form und Inhalt der Datenbank lieferten. Die Resultate der Untersuchung definitorischer Elemente finden sich in Abschnitt 3.2.2. Sie bieten einen Einblick in mögliche Fehlerquellen und lassen Aussagen zur Nützlichkeit von Definitionsfeldern zu.

3.2.1 Konsistenz und Kohärenz

Die Vorbereitungsarbeiten ermöglichen einen Rückschluss auf den Zweck der SIMATIC-Datenbank und einige Beobachtungen zu Konsistenz und Kohärenz von Datenbankinhalt und -struktur. Eine Übersicht der für die nachfolgenden Bemerkungen relevanten Kennzahlen ist in Tabelle 2 dargestellt.

[112] Schneider 1997, Microsoft Press 2001, Winkler 2000, IBM 1985

Tabelle 2 Auszug aus der für die SIMATIC-TermDB erstellten Statistik

Feld	Häufigkeit
Einträge insgesamt	7'379
italienische Benennungen	8'685
deutsche Benennungen	8'585
spanische Benennungen	7'749
französische Benennungen	6'639
englische Benennungen	6'093
Benennungsquellen	2'248
Kontexte	4'343
Definitionen	2'361

Die Anzahl mehrsprachiger Einträge kann aufgrund der in Tabelle 2 dargestellten Häufigkeit der jeweiligen Sprachfelder als hoch eingeschätzt werden. Insgesamt handelt es sich um 7379 Einträge. Die Anzahl deutscher, englischer, französischer, englischer und spanischer Benennungen liegt jeweils zwischen rund 6100 und 8700 Vorkommen. Mehrsprachige Einträge sind demzufolge häufig. Das zeigt, dass die Datenbank in erster Linie für die Festlegung anderssprachiger Benennungen genutzt wird und die Erklärung von Begriffen zweitrangig ist. Dieser Verwendungszweck spiegelt sich im unter Abschnitt 3.1.2 dargestellten Minimaleintrag. Neben den automatisch generierten Feldern (Eintragsnummer, Felder mit Änderungsdatum und -benutzer) weist jeder Eintrag Benennungsfelder, zwei die Benennung modifizierende Attributfelder und eine Sachgebietsangabe auf. Weiterführende Informationen finden sich meist in Form von Kontexten (4343 Vorkommen), was wie die Häufigkeit mehrsprachiger Einträge ein Hinweis auf die übersetzungsorientierte Ausrichtung der Datenbank ist. Definitionen (2361 Instanzen) erscheinen nur für ungefähr ein Drittel der Begriffe.

Die Statistik der SIMATIC-TermDB erlaubt eine ungefähre Beurteilung des Datenbankinhalts. Als Erstes zeigte sich die mangelnde Häufigkeit von Quellenangaben. Herkunftsangaben (2248) sind für ungefähr ein Drittel der Benennungen vorhanden. Bei zwei Dritteln der Begriffe ist also der Ursprung nicht belegt. Aus terminologischer Sicht sind Quellenangaben unabdingbar, da sie weitere Recherchen im Umfeld der erfassten Daten erleichtern und Auskunft über die Zuverlässigkeit des Eintrags geben.[113] Quellen können auch Rückschlüsse auf das Fachgebiet, aus dem die Daten stammen, ermöglichen und in

[113] KÜWES 1990, Kap. 4, S. 2

dieser Form als definitorisches Element wirken. Die systematische Anführung der Benennungsquellen ist entscheidend für die Qualität der erfassten Daten.

Bei der Ausfilterung vordefinierter Angaben zeigte sich eine mangelnde Kohärenz bei der Auswahl der Sachgebiete. Wie in Tabelle 3 dargestellt, figurieren sieben von insgesamt 13 Fachgebieten zwischen ein- und dreissigmal. Zwei Fachgebieten (*SIMATIC S7* und *Allgemeine Elektrotechnik*) wurden demgegenüber über 2'500 Einträge zugeordnet. Bei den Teilgebieten zeigte sich ein ähnliches Bild. Von gesamthaft 13 Teilgebieten weisen fünf eine Häufigkeit von unter 10 auf. Da das Fachgebiet den extralinguistischen Bezugspunkt der Definition verdeutlicht, ist eine sinnvolle Gliederung für die Klassifikation und damit das Verständnis eines Begriffs relevant. Im vorliegenden Fall ist zu überprüfen, ob einerseits wenig gebrauchte Fach- und Teilgebiete in andere überführt werden können und andererseits eine Erweiterung der zur Auswahl stehenden Gebiete zur genaueren Identifikation der Einträge beitragen kann.

Tabelle 3 Inhalte des Felds "Fachgebiet" in der SIMATIC-TermDB

Feldinhalt	**Häufigkeit**
SIMATIC S7	3532
Allgemeine Elektrotechnik	2796
SIMATIC Kommunikation	465
SIMATIC HMI	285
SIMATIC PCS7	176
SIMATIC Allgemeiner Begriff	78
CD Power Distribution Products	30
SIMATIC S7-200/STEP 7-Micro	28
CD Control Products	11
SIMATIC Technologie	4
SIMATIC Eigenname	3
CD Gesamt	2
Norm IEC 60050: IEV Kap. 301 – General Terms on Measurements in Electricity	2

Die Statistik zeigte drei formale Schwächen der Datenbank. Der erste Punkt betrifft das Feld "Abkürzung/Vollbegriff". Die SIMATIC-TermDB verwendet das Feld "Abkürzung/Vollbegriff" sowohl für Vollbenennungen als auch für Kurzformen. Vermutlich enthält dieses Feld die weniger gebräuchliche Form, während in die Sprachfelder die hauptsächlich verwendete Benennung geschrieben wird. Die Erfassung der Voll- und Kurzformen im jeweiligen Sprachfeld erleichtert die Abfrage, da die Suche sich dann nur über

ein Feld erstreckt. Dazu kann ein Attributfeld mit zwei vordefinierten Elementen (Kurz- und Vollbenennung) angefügt werden, um Kurzformen als solche zu kennzeichnen.
Ein zweiter Punkt ist die Darstellung von Synonymen, Schreibvarianten oder Zusatzinformationen im Benennungsfeld mit Hilfe von Sonderzeichen. Wie vorgängig besprochen ist die Erfassung nicht einheitlich. Klammern werden bei Synonymen, Schreibvarianten oder Zusatzinformationen benützt. Synonyme können auch durch Kommata getrennt sein. Das Komma wird zudem häufig für Substantiv-Adjektiv-Konstruktionen verwendet. Eine einheitliche Darstellung von Synonymen und Mehrwortbenennungen erleichtert die Abfrage. Wenn Synonyme einzeln im jeweiligen Sprachfeld erfasst werden, können sie direkt (d.h. ohne unscharfe Suche) angewählt werden. Dasselbe gilt für die konsistente Erfassung von Mehrwortbenennungen und Fachwendungen.
Schliesslich zeigte die Ausfilterung des Grammatikfeld-Inhalts, dass es nicht als Auswahlfeld konzipiert ist. Die grammatikalischen Angaben können frei formuliert werden und sind unterschiedlich gehalten. Konsistenz ist jedoch anzustreben; daher sollten die Feldwerte in einer Auswahlliste vordefiniert werden.

3.2.2 Definitionen

3.2.2.1 Fehlende Definitionen

Der Vergleich der Datenbestände gestaltete sich bei Einträgen ohne Definition schwierig. Das unter 3.1.4 vorgestellte Bewertungsschema half bei der Beurteilung solcher Fälle, konnte jedoch keine definitiven Aussagen ermöglichen. Die manuelle Prüfung zeigte weiter, dass verschiedentlich jeweils zwei Einträge innerhalb der SIMATIC-TermDB grosse Ähnlichkeit miteinander aufwiesen. In diesen Eintragspaaren stimmten mehrere Benennungen miteinander überein, ohne dass klar ersichtliche Bedeutungsunterschiede vorhanden waren.

<Eintragsnummer>2546
<Fachgebiet>Allgemeine Elektrotechnik
<ITA>elemento di ritardo
<Abgestimmt>ja
<Status>frei
<Last updated by>6RE
<Benennungs-Quelle>Ing. Ghizzoni
<ENG>delay element
<Abgestimmt>ja
<Status>frei
<Last updated by>6RA
<Kontext>higher-order delay element
<FRA>opérateur à retard
<Abgestimmt>ja
<Status>frei
<Last updated by>6JM
<GER>Verzögerungsglied
<Abgestimmt>ja
<Status>frei
<Last updated by>6RE
<Last updated>11.10.2000
<SPA>elemento de retardo
<Abgestimmt>ja
<Status>frei
<Last updated by>6VA

<Eintragsnummer>41646
<Fachgebiet>SIMATIC S7
<ITA>elemento di ritardo
<Abgestimmt>ja
<Status>gesperrt
<Last updated by>TDB
<Last updated>07.04.99
<ENG>delay element
<Abgestimmt>ja
<Status>bevorzugt
<Last updated by>TDB
<Last updated>06.04.99
<ENG>monoflop
<Abgestimmt>ja
<Status>gesperrt
<Bemerkung>monoflop war ursprünglich als Synonym eingetragen.
<GER>Verzögerungsglied
<Abgestimmt>ja
<Status>bevorzugt
<Last updated by>TDB
<Last updated>06.04.99
<Bemerkung>vergl. ^Einschaltverzögerungsglied^, ^Ausschaltverzögerungsglied^
<SPA>elemento de retardo
<Abgestimmt>ja
<Status>bevorzugt
<Last updated by>TDB
<Last updated>07.04.99
<SPA>monoestable
<Abgestimmt>ja
<Status>gesperrt
<Bemerkung>monoestable war ursprünglich als Synonym eingetragen.

Die beiden Einträge weisen übereinstimmende italienische (*elemento di ritardo*), englische (*delay element*), deutsche (*Verzögerungsglied*) und spanische (*elemento de retardo*) Benennungen auf. Im Eintrag auf der rechten Seite (Eintragsnummer 41646) ist keine französische Entsprechung vorhanden, jedoch als veraltet gekennzeichnete spanische und englische Synonyme (*monoflop* bzw. *monoestable*). Die zwei Einträge weisen unterschiedliche Fachgebietsangaben auf (*Allgemeine Elektrotechnik* und *SIMATIC S7*). Ohne weiterführende Angaben, insbesondere ohne Definition, lässt sich schlecht bestimmen, ob die zwei Einträge tatsächlich zwei Begriffe darstellen oder ob es sich um Dubletten handelt, also redundante Einträge mit demselben Referenten. Die Fachgebietsangabe ist der einzige Hinweis darauf, dass ein Verzögerungsglied in einem SIMATIC S7-System möglicher-

weise ein anderes Konzept darstellt, als das in der allgemeinen Elektrotechnik der Fall ist. Diese Vorkommen wurden markiert, konnten jedoch nicht beurteilt werden.

3.2.2.2 Qualität der Definitionen

Da wenige übereinstimmende Deutschbenennungen in den zwei Beständen vorhanden waren, konnte die Qualität der Definitionen nur anhand von 71 Einträgen mit 74 unterschiedlichen Definitionen (drei Einträge enthielten zwei inhaltlich verschiedene Definitionen) geprüft werden. Aus Gründen der Praktikabilität wurden die Definitionen als in sich geschlossene Einheiten betrachtet. Der Bezug zum Begriffssystem wurde nicht beurteilt, da die SIMATIC-TermDB Begriffe unterschiedlicher Fachrichtungen enthält. Die Sachgebietsauswahl in der Datenbank zeigt, dass Begriffe aus der allgemeinen Elektrotechnik, aber auch aus dem Hard- oder Softwarebereich erfasst sind. Die mangelnde Kohärenz bei der Fachgebietszuordnung der Einträge erschwert die Abgrenzung von Begriffsfeldern zusätzlich. So wurden die in Abschnitt 2.1.3.1 diskutierten Anforderungen, dass alle Begriffe unter einem sachlichen Gesichtspunkt zu definieren und die als Bezugspunkte verwendeten Benennungen demselben Begriffssystem zu entnehmen sind, nicht untersucht. Aufgrund der fehlenden Begriffsfelder kam auch die Beurteilung unterscheidender Merkmale bei Inhaltsdefinitionen nicht zum Tragen.

3.2.2.2.1 Qualität inhaltlicher Art

Von den insgesamt 74 geprüften Definitionen wurden 53 als inhaltlich ausreichend bewertet. Formfehler wurden angemerkt und werden nachfolgend diskutiert. 35 davon hatten die Form einer Inhaltsdefinition, 17 wurden als *Begriffsumschreibungen* nach KÜWES eingestuft.[114] Dazu fand sich eine Umfangsdefinition. 21 der geprüften Definitionen liessen inhaltliche Mängel erkennen. Die Qualitätsdefizite konnten in drei Gruppen aufgeteilt werden: begriffsbezogene Mängel, falsch platzierte Feldinhalte und Mängel im eigentlichen Definitionsinhalt.

Zur ersten Kategorie gehören fünf Fälle, in denen ein Eintrag mehrere Begriffe enthält. In Beispiel (14) geht aus der Definition selbst hervor, dass mehrere Begriffe aufgeführt werden. Terminologiedatenbanken orientieren sich nicht an Benennungen, sondern an den durch sie referenzierten Gegenständen. Ihre Struktur ist damit konzeptbezogen und lässt

[114] KÜWES 1990, Kap. 4, S. 5

keine begriffliche Durchmischung zu. Eine weitere Form begriffsbezogener Mängel zeigte sich bei zwei Einträgen. Sie weisen eine Definition auf, die nicht dem beschriebenen Gegenstand entspricht. In Beispiel (15) ist aufgrund des englischen Synonyms *re-chargeable battery* zu vermuten, dass in diesem Eintrag *AKKU* als 'Stromspeicher' und nicht als 'Bestandteil der CPU' beschrieben wird.

(14) *Betriebsart: Unter Betriebsart verstehen wir 1. die Anwahl eines Betriebszustandes der CPU mit dem Betriebsartenschalter oder mit dem PG 2. die Art des Programmablaufs in der CPU.*

(15) *AKKU (*englisch: *ACCU*, mit Synonymeintrag *rechargeable battery): Akkumulatoren sind Register in der CPU und dienen als Zwischenspeicher für Lade-, Transfer- sowie Vergleichs-, Rechen- und Umwandlungsoperationen.*

Neun Inhalte von Definitionsfeldern entsprechen der zweiten Mängelkategorie, den falsch platzierten Feldinhalten. Dabei handelt es sich um Text, der entweder den Charakter einer Bemerkung (16), eines Beispiels (17), eines Kontexts (18) oder einer Vollbenennung (19) hat.

(16) *Systemparameter: Parameter im COM ET 200.*

(17) *Adapterkabel: Ein Adapterkabel wird z. B. zum Anschluss des PGs oder des ET 200-Handhelds an ET 200C benötigt. Dieses Adapterkabel hat einen 9-poligen Busanschlussstecker in Schutzart IP 20 und einen 12-poligen Busanschlussstecker in Schutzart IP 66/67.*

(18) *auflösen: Eine Gruppe aus Einzelobjekten wird aufgelöst.*

(19) *EGB-Richtlinien: Richtlinien für elektrostatische gefährdete Bauelemente*

In Beispiel (16) kann die Information, dass es sich um einen Parameter in einem bestimmten System handelt, nützlich sein. Sie genügt aber nicht als Erläuterung des Wesens eines Systemparameters. Beispiel (17) beschreibt eine mögliche Ausführung eines Adapterkabels und führt einen Anwendungsfall an. Diese Angaben haben Beispielcharakter und gehören ins Bemerkungsfeld. In Beispiel (18) wird der Begriff *auflösen* im Kontext verwendet. Beispiel (19) zeigt eine ausgeschriebene Vollbenennung im Definitionsfeld. Sie sollte vorzugsweise ins Benennungsfeld geschrieben werden, da die Vollbenennung keine inhaltliche Definition des Begriffs darstellen kann.

Als dritte Kategorie fanden sich Mängel, die sich auf den Definitionsinhalt bezogen. Vier Definitionen weisen eindeutige Zirkelbezüge auf. Da Zirkelhaftigkeit bei Komposita häufig schwierig zu umgehen ist, beschränkte sich die Evaluation auf die Identifizierung eindeutiger Zirkelbezüge, d.h. beide Teile einer Mehrwortbenennung oder eines Kompositums figurieren in der Definition. In Beispiel (20) wird durch die Wiederholung der

beiden Wortkomponenten weder das Wesen von Übergängen noch dasjenige von Netzen erklärt und nur wenig neue Information vermittelt.

(20) *Netzübergang: Übergang zwischen Subnetzen eines Gesamtnetzes.*

Bei Grenzfällen wurde in der Auswertung eine entsprechende Anmerkung angebracht. So wurde Beispiel (21) nicht als eindeutige Zirkeldefinition gewertet, da nur der Kern des Kompositums in der Definition wiederholt wird. Entsprechend wurde nur in einer Anmerkung darauf hingewiesen, dass *Quittierung* vorzugsweise zu ersetzen ist, bspw. durch *Empfangsbestätigung*:

(21) *Einzelquittierung: Im Gegensatz zur Sammelquittierung wird bei der Einzelquittierung nur jeweils eine Meldung quittiert.*

3.2.2.2.2 Qualität formaler Art

Ebenfalls als Anmerkungen wurden Formfehler notiert. Sie beziehen sich grösstenteils auf unnötige Definitionsteile. Zahlreiche Definitionen enthalten die definierte Benennung, deren Wiederholung überflüssig ist und die Definition unnötig verlängert, wie das in Beispiel (22) der Fall ist:

(22) *Buskabel: Das Buskabel ist ein Kabel, mit dem Busteilnehmer verbunden werden.*

In 13 Fällen enthalten Definitionen Zusatzinformationen, die vorzugsweise im Bemerkungsfeld erfasst werden. Die letzten beiden Sätze in Beispiel (23) enthalten für die Definition nicht zentrale Elemente und wären daher im Bemerkungsfeld besser platziert.

(23) *Funktionsplan: Der FUP (Funktionsplan) ist eine der drei Programmiersprachen von STEP5 und STEP7. FUP benutzt zur Darstellung der Logik die von der Bool'schen Algebra bekannten logischen Kästchen. Ausserdem können komplexe Funktionen (z.B. mathematische Funktionen) direkt in Verbindung mit den logischen Kästchen dargestellt werden. Eine Übersetzung in andere Programmiersprache (z.B. Kontaktplan) ist möglich.*

3.3 Zusammenfassung

Das Siemens-Projekt ermöglichte Beobachtungen zur systematischen Evaluation einer Terminologiedatenbank sowie zur Nützlichkeit terminologischer Angaben und zur Realisierbarkeit theoretischer Anforderungen an einen Datenbestand. Eine rein statistische Auswertung der Datenbankfelder lieferte bereits aussagekräftige Anhaltspunkte zur inhaltlichen und formalen Beurteilung der Datenbank. Sie zeigte, dass die Einteilung der Fach- und Teilgebiete nicht kohärent ist. Einige Gebiete erscheinen nur vereinzelt, andere werden mit einer so grossen Häufigkeit verwendet, dass sie kein aussagekräftiges Unter-

scheidungsmerkmal gegenüber andern Begriffen mehr sein können. Weiter sind Benennungsquellen, die Auskunft über Herkunft und Zuverlässigkeit der Daten geben, in weniger als einem Drittel der Einträge angeführt. Die Statistik ermöglichte auch einige Aussagen zur Struktur der Datenbank. Das Feld "Abkürzung/Vollbegriff" enthält sowohl Kurzformen als auch Vollbenennungen. Die Suche nach Benennungen würde durch eine Erfassung von Abkürzungen und Vollformen im jeweiligen Sprachfeld vereinfacht und zielgerichteter gestaltet. Die konsistente Erfassung von Synonymen und Fachwendungen erleichtert die Suche weiter. Synonyme und Wendungen waren teilweise mit Hilfe von Kommata dargestellt, aber auch einzeln erfasst worden.

Die Nützlichkeit terminologischer Felder trat im Rahmen der manuellen Überprüfung zutage. Einträge ohne Definition liessen sich kaum mit dem genormten Bestand vergleichen, da der Begriff schlecht fassbar ist. Zudem zeigte sich, dass bei sehr ähnlichen Einträgen eine Beurteilung ihrer Relevanz bzw. Redundanz ohne Definition schwierig ist. Entsprechend lassen sich auch Synonyme nicht bestimmen oder Benennungen revidieren. Bereinigungsarbeiten werden allgemein erschwert. In solchen Fällen ist häufig die Sachgebietsangabe der einzige Hinweis auf eine unterschiedliche Herkunft oder Verwendung, was die Wichtigkeit einer kohärenten Gebietsunterteilung unterstreicht. Nützlich wären auch Quellenangaben gewesen, da sie den Ursprung der Einträge klären.
Die Überprüfung zeigte weiter, dass die in der Literatur formulierten Anforderungen an Definitionen bei der nicht systematischen Terminologiearbeit nur teilweise anwendbar sind. Ein ausgearbeitetes und abgegrenztes Begriffssystem ist bei der übersetzungsorientierten Arbeit selten gegeben, und damit in Zusammenhang stehende Bedingungen (z.B. Wahl der Bezugspunkte aus demselben Begriffssystem oder Wesentlichkeit der unterscheidenden Merkmale bei der Inhaltsdefinition) können nicht erfüllt werden.

Auf der Basis der Umfrage bei Übersetzern eines Sprachendienstes und den im Rahmen eines Terminologieprojekts gemachten Erfahrungen kann die Definition auch bei der übersetzungsorientierten Arbeit als relevante Datenkategorie eingestuft werden. Die Erstellung von Definitionen ist jedoch aus Zeitgründen häufig nicht möglich, da der Schwerpunkt bei der Auffindung zielsprachlicher Äquivalente liegt. Zudem gestaltet sich die laufende Anpassung von Definitionen an sich verändernde Begriffsinhalte schwierig. Als mögliche Lösung bietet sich die automatische Extraktion von Definitionskandidaten an. Sie mindert

den Aufwand, mit dem definitionsähnliche Sätze gefunden werden können. Aus diesem Grund soll in Kapitel 4 die automatische Identifikation von Definitionskandidaten erörtert werden.

4 EXTRAKTION VON DEFINITIONSKANDIDATEN

Kapitel 4 beschäftigt sich mit der automatischen Identifikation kontextualisierter Definitionen. Ziel ist zum einen die Erörterung bereits implementierter Ansätze zur Erkennung von definitorischen Sätzen und zum andern die Entwicklung von Kriterien zur Extraktion von Definitionskandidaten aus einem deutschsprachigen Korpus. Als Erstes werden anhand zweier empirischer Studien die Erscheinungsformen von in Texten eingebetteten definitorischen Sätzen besprochen. Diese beiden Arbeiten vermitteln durch Einzelbeobachtungen einerseits und die systematische Auswertung eines Korpus andererseits einen Eindruck der Arten kontextualisierter Begriffsbestimmungen. In Kapitel 4.2 folgt eine Erörterung der Ansätze zur automatisierten Extraktion von definitorischen Sätzen. Pearson (Abschnitt 4.2.1) bietet eine ausführliche Darstellung der Identifikation verschiedener definitorischer Formen.[115] Die in den Abschnitten 4.2.2 und 4.2.3 vorgestellten Arbeiten von Meyer und Rebeyrolle gehen von der Suche nach Beziehungen zwischen Begriffen aus und erweitern die verwendeten Muster, um definitorische Aussagen zu finden.[116] Diese drei Arbeiten beschäftigen sich mit der automatischen Erkennung von definitorischen Sätzen an sich. Zwei weitere Studien setzen den Schwerpunkt auf die Gewinnung und Repräsentation der Segmente einer Definition. Die Verfahren von Bowden et al. und Büchel et al. (Abschnitte 4.2.4 und 4.2.5) bezwecken neben der Identifikation von Definitionen die Darstellung ihrer Komponenten in strukturierter Form.[117] Weitere Arbeiten befassen sich im Rahmen der Wissensakquisition mit der automatischen oder halbautomatischen Identifikation von Begriffsbeziehungen. Wie in Abschnitt 2.1.1 erörtert wurde, basieren die Inhalts- und die Umfangsdefinition auf den Beziehungen zwischen Begriffen, d.h. sie verdeutlichen die hierarchischen Beziehungen in einem Begriffssystem. Im Fall der Inhaltsdefinition ist dies die Hyponymie-, bei der Umfangsdefinition die Meronymie-Beziehung. Daher ist der Erkennung definitorischer Sätze und der Identifikation von Begriffsbeziehungen dieselbe Basis gemein. Aus diesem Grund wird in Kapitel 4.3 eine Auswahl der Untersuchungen zur Erkennung von Begriffsbeziehungen vorgestellt. Die in Kapitel 4.2 dargestellten Arbeiten zur Extraktion von Definitionskandidaten bilden den

[115] Pearson 1998
[116] Meyer 2001; Rebeyrolle 2000
[117] Bowden et al. 1996; Büchel et al. 1995

Ausgangspunkt für die in Kapitel 4.4 erörterte Implementierung von Kriterien zur Bestimmung definitorischer Sätze in einem deutschsprachigen Testkorpus.

4.1 Empirische Untersuchungen zur Form von Definitionen

Eine Reihe von Arbeiten befasst sich mit den Erscheinungsformen kontextualisierter Definitionen. Nachfolgend werden zwei Artikel erörtert, deren Untersuchungen Anhaltspunkte für die automatische Extraktion von definitorischen Sätzen aus Texten bieten. Drei weitere Arbeiten beschäftigen sich anhand von korpusbasierten Studien mit Begriffsbestimmungen. Lambrou basiert seine Arbeit auf einer Auswahl von wissenschaftlichen Unterrichtsbüchern und untersucht die Anordnung der Elemente, die eine Definition konstituieren.[118] Bramki et al diskutieren anhand einiger Kapitel eines Wirtschaftslehrbuchs die Unterteilung von Definitionen in verschiedene Typen.[119] Chaudron befasst sich mit gesprochener Sprache.[120] Er untersucht die Merkmale von Definitionen unter verschiedenen Gesichtspunkten; so erörtert er beispielsweise ihre morphologischen und syntaktischen Eigenheiten.

4.1.1 Trimble

Trimble beschäftigt sich mit der Form der englischen Sprache im wissenschaftlichen und technischen Bereich (*English of science and technology (EST)*).[121] Seine Aussagen zu Definitionen basieren nicht auf der systematischen Auswertung eines Korpus, sondern auf Einzelbeobachtungen in wissenschaftlichen und technischen Dokumenten. Trimble entwirft ein Klassifikationsschema für definitorische Sätze, macht allerdings keine Angaben zur Häufigkeit der Typen oder ihrer genauen Struktur. So finden sich beispielsweise keine Anhaltspunkte zur Kategorie von Verben, die Ober- und Unterbegriff verbinden, oder zu typischen syntaktischen Mustern.

Trimbles grundsätzliche Unterscheidung ist zwischen einfachen und komplexen Definitionen (*simple and complex definitions*). Einfache Definitionen werden in einem Satz aus-

[118] Lambrou 1979, zitiert nach: Flowerdew 1992, S. 205
[119] Bramki et al. 1984
[120] Chaudron 1982, zitiert nach: Flowerdew 1992, S. 206
[121] Trimble 1985

gedrückt, komplexe Definitionen erstrecken sich über mehrere Sätze.[122] Komplexe Definitionen werden als Erweiterungen einfacher Definitionen betrachtet: "Characteristically, most expanded definitions are developed in paragraph units and have, as a rule, a simple definition — formal or semi-formal — for their core statement."[123] Sie werden hier nicht näher erörtert, da der Schwerpunkt der vorliegenden Arbeit auf definitorischen Aussagen in Einzelsätzen liegt. Die einfache Definition gliedert er in drei Untergruppen: formale, semiformale und nicht formale Definitionen (*formal*, *semi-formal* und *non-formal*).

Die formale Definition entspricht der Inhaltsdefinition: "[The formal definition] is, of course, the well-known equation-like '*Species* = *Genus* + *Differentia*', usually called 'formal' because of its rigidity of form."[124] Sie gibt also den zu definierenden Begriff, sein Hyperonym und die ihn von anderen Begriffen unterscheidenden Merkmale an.

Im Gegensatz zur formalen Definition enthält die semiformale Begriffsbestimmung nur zwei Angaben, den Begriff selbst und die ihn unterscheidenden Merkmale: "By definition, a semi-formal definition contains only two of the three basic defining elements: the term being defined and the statement of differences."[125] Das Hyperonym fehlt, möglicherweise weil es als bekannt vorausgesetzt oder als nicht relevant betrachtet wird:

(24) *An anemometer registers the speed of the wind on a dial or gage*.[126]

In Beispiel (24) fehlt das Hyperonym (bspw. in der Form von *is a meteorological instrument*) und der Begriff wird durch seine Funktion beschrieben.

Nicht formale Definitionen geben neben dem Begriff ein Wort oder eine Phrase an, die den Sinn des Begriffs erläutern oder ein herausragendes Merkmal bezeichnen: "The function of a non-formal definition is to define in a general sense so that a reader can see the familiar element in whatever the new term may be. [...] Most non-formal definitions are found in the form of synoyms."[127] Als Beispiel führt Trimble die Erläuterung *A helix is a spiral* an, in der ein typisches Merkmal als definitorisches Element verwendet wird.[128] Zu den nicht formalen Definitionen zählt Trimble auch Negativdefinitionen und Erläuterungen durch Angabe von Synonymen oder Antonymen.

---

[122] Trimble 1985, S. 75
[123] Trimble 1985, S. 81
[124] Trimble 1985, S. 75 f.
[125] Trimble 1985, S. 77
[126] Trimble 1985, S. 80
[127] Trimble 1985, S. 78
[128] Trimble 1985, S. 78

Trimbles Klassifikation einfacher Definitionen basiert auf der Vollständigkeit der angegebenen Information. Er konzentriert sich auf der Inhaltsdefinition ähnliche Arten und erwähnt keine anderen Formen. Wie eingangs angemerkt geht er nicht auf die formalen Merkmale verschiedener Definitionsarten ein. Sein Schema ist jedoch für die automatische Extraktion von definitorischen Sätzen hilfreich, da er die semantische Grundstruktur von Begriffsbestimmungen beschreibt.

4.1.2 Flowerdew

Im Gegensatz zu Trimble basiert Flowerdew seine Untersuchung auf der Auswertung eines Testkorpus. Es setzt sich aus einer Reihe naturwissenschaftlicher Vorlesungen von englischsprachigen Dozenten für fremdsprachige Studenten zusammen.[129] Er prüfte 329 Definitionen, die 315 Begriffe erläutern.[130] Flowerdew befasst sich ausschliesslich mit gesprochener Sprache. Er untersucht unter anderem die Merkmale verschiedener Definitionstypen und ihre sprachliche Kennzeichnung.

Flowerdews Klassifikation von Definitionsarten basiert auf Trimbles Schema. Er unterscheidet drei Haupttypen von Definitionen: formale und semiformale Definitionen sowie Substitution (*substitution*).[131] Für die formale und semiformale Definition übernimmt er Trimbles Begriffsbestimmungen. Bei der Substitution wird nach Flowerdew die Benennung durch ein Wort, Wortteil oder eine Phrase mit ähnlicher Bedeutung ersetzt. Sie entspricht also zu einem grossen Teil Trimbles nicht formaler Definition. Flowerdew unterscheidet drei Arten von Substitution. Die Benennung kann durch ein Synonym (z.B. *by fuse I mean join together*), eine Paraphrase (z.B. *electropositive is likes to become positively charged*) oder eine Derivation ersetzt werden. Bei der Derivation wird nur ein Teil einer Benennung erklärt, aus dem sich dann aber der Begriff ableiten lässt, bspw. *the pyliferous layer is so called because it's hairy*.[132] Flowerdew identifiziert einen definitorischen Nebentypen, die Ostension. Bei ihr wird auf eine Visualisierung des Begriffs (bspw. eine Fotografie oder ein Diagramm) verwiesen.[133]

[129] Flowerdew 1992a, S. 203
[130] Flowerdew 1992a, S. 207
[131] Flowerdew 1992a, S. 209 ff.
[132] Flowerdew 1992a, S. 211
[133] Flowerdew 1992a, S. 212

Flowerdews Untersuchung der syntaktischen Strukturen der drei Haupttypen zeigte einige typische Muster. Die formale Definition besteht üblicherweise aus Nominalphrase – Kopula – Nominalphrase. Dazu kommen ein Relativsatz oder andere prä- oder post-modifizierende (Beispiel 25) Elemente, zu deren Formen er keine näheren Angaben macht.[134]

(25) *[…] so nephridiopores are very tiny holes lying on the ventral surface of the earth worm*[135]

Die typische Struktur einer semiformalen Definition umfasst nach Flowerdew ebenfalls Nominalphrase – Kopula – Nominalphrase, aber ohne einen nachfolgenden Relativsatz. Er führt allerdings keine Beispiele dieser Struktur an. Fraglich ist, ob es sich tatsächlich meist um Kopula handelt oder ob bei der semiformalen Definition auch andere Verben gebräuchlich sind, wie dies in Beispiel (24) von Trimble der Fall ist (*an anemometer registers […]*) Die Form einer Substitution hat Ähnlichkeit mit derjenigen der semi-formalen Definition. Bei der Substitution sind die beiden Nominalphrasen jedoch häufig nicht durch eine Kopula verbunden, sondern sie stehen in Apposition, die entweder explizit gekennzeichnet ist (z.B. durch *or*) oder durch die Intonation signalisiert wird.

(26) *[…] by hydrolysis or the addition of water […]*[136]

In formalen und semiformalen Definitionen erscheint der Begriff nicht notwendigerweise am Satzanfang. Nach dem Prinzip des Endfokus wird das betonte semantische Element am Ende des Satzes platziert. Ist der Begriff bereits eingeführt worden, stellt er die bekannte Information dar und erscheint am Satzanfang. Wenn er jedoch als neues Element vorgestellt wird, ändert die semantische Ordnung der Definition, und er erscheint am Satzende:

> "Where the term has not been previously established, the semantic ordering of the definition is reversed, with the term coming at the end, in so-called nominal definition, e.g. *on the ventral surface of the earthworm there are small projections which are known as the chaetae.*"[137]

Definitionen können durch syntaktische oder lexikalische Elemente signalisiert werden. Kopulative Strukturen bilden das häufigste syntaktische Element in Flowerdews Korpus; die Art der auftretenden Verbindungsverben wird allerdings nicht spezifiziert. Als zweites

[134] Flowerdew 1992b, S. 171
[135] Flowerdew 1992a, S. 210
[136] Flowerdew 1992b, S. 171
[137] Flowerdew 1992b, S. 168

syntaktisches Element nennt Flowerdew Relativsätze, die sich bei formalen Definitionen finden. Ungefähr die Hälfte der Definitionen in Flowerdews Korpus ist lexikalisch gekennzeichnet. Bei den lexikalischen Merkmalen unterscheidet Flowerdew zwischen *boosters* und *downtoners*. *Boosters* definiert er als "linguistic (in this case lexical) items that signal clearly the illocutionary force of a speech act".[138] Sie kennzeichnen also Definitionen explizit als solche. Beispiele sind Ausdrücke mit dem Verb *to call* (z.B. *we call, is/are called*), die in Flowerdews Korpus häufig auftraten. Andere Formen sind *or, known as* oder *that is*.[139] Im Gegensatz zu *boosters* können *downtoners* die Aussagekraft einer Definition mindern: "The force of a definition can be downgraded by the use of a downtoner. Downtoners may be realized as adverbials, modal *can*, and non-factive predicators (expressions which do not commit a speaker to the truth of a proposition)."[140] Als Beispiele für den Gebrauch von *downtoners* führt Flowerdew folgende Sätze an:

(27) *We **can** say the plasma membrane is the membrane around the cell.* (Modalverb)
(28) ***A way of defining** a metal is **by saying** that it is an element that readily forms cations* (*non-factive predicator*)[141]

Flowerdew befasst sich wie Trimble ausschliesslich mit der Inhaltsdefinition und ihr verwandten Formen. Im Gegensatz zu Trimble beschäftigt er sich jedoch mit ihrer syntaktischen Form, der syntaktisch-lexikalischen Signalisierung und der Anordnung der eine Definition konstituierenden semantischen Elemente. Flowerdews korpusbasierte Beobachtungen bieten nützliche Hinweise zur Form kontextualisierter Definitionen und damit zu ihrer automatischen Erkennung. Die in Abschnitt 4.2.1 erörterte Arbeit von Pearson setzt die von Trimble und Flowerdew gemachten Beobachtungen für die automatische Erkennung von Begriffsbestimmungen ein.

[138] Flowerdew 1992b, S. 172
[139] Flowerdew 1992a, S. 212
[140] Flowerdew 1992b, S. 172
[141] Flowerdew 1992b, S. 173

4.2 Ansätze zur Extraktion von Definitionen und definitorischen Kontexten

4.2.1 Pearson

Pearsons Arbeit ist sehr umfangreich und bot zahlreiche Anhaltspunkte für die in Kapitel 4.4 beschriebene Extraktion deutscher Definitionskandidaten. Aus diesem Grund werden die darin spezifizierten Muster für die Identifikation definitorischer Sätze nachfolgend ausführlich erörtert. Bei der Extraktion von definitorischen Erläuterungen unterscheidet Pearson zwischen *simple formal defining expositives, complex formal defining expositives, semi-formal defining expositives* und *dictionary type definitions*.[142] Diese Kategorisierung basiert zum einen auf Austin[143], zum andern auf den Untersuchungen von Trimble und Flowerdew. Gemäss Austin werden Expositiva (*expositives*) verwendet "in acts of exposition involving the expounding of views, the conducting of arguments, and the clarifying of usages and of references".[144] Pearson bezeichnet damit Definitionen, die nicht zum ersten Mal formuliert, sondern vom Autor oder Sprecher wiederholt werden: "The definition is not being formulated for the first time and consequently, the authors are in fact reporting rather than doing."[145] Pearsons *simple formal defining expositives* und *complex formal defining expositives* kommen der einfachen bzw. komplexen formalen Definition von Trimble gleich, also der Definition eines Begriffs durch Hyperonym und unterscheidende Merkmale in einem oder mehreren Sätzen. Desgleichen finden die *semi-formal defining expositives* ihre Entsprechung in Trimbles semiformaler Definition, in der sich ein Begriff durch ein charakteristisches Merkmal erklärt.[146] Unter *dictionary type definitions* versteht Pearson glossarartige Einträge, die sich in den Korpora fanden.[147] Dieser letzte Typ wird nachfolgend nicht behandelt, da es sich dabei um typographisch explizit als Definitionen gekennzeichnete Sätze handelt. Diese Art von Äusserungen ist für die Extraktion kontextualisierter definitorischer Sätze nicht relevant.

---

[142] Pearson 1998, S. 135 ff.
[143] Austin 1962
[144] Austin 1962, S. 161, zitiert nach: Pearson 1998, S. 116
[145] Pearson 1998, S. 117
[146] Pearson 1998, S. 135
[147] Pearson 1998, S. 136

Die Unterscheidung nach Austins Sprechakten ist für die automatische Definitionserkennung wenig bedeutsam. Ableiten lässt sich daraus, dass Expositiva in informationsvermittelnden Textsorten häufig erscheinen. Entsprechend basiert Pearson ihre Studie auf drei Korpora, deren Texte die Vermittlung von Wissen zum Ziel haben.[148]

Pearsons Verfahren liegt eine vorgängige Termerkennung zugrunde. Die dazu verwendeten Termbildungsmuster wurden von Pearson spezifiziert und je nach Korpus angepasst. Die Kriterien für die Suche nach Begriffen sind sehr restriktiv formuliert, da die Termerkennung nur ein Vorbereitungsschritt zur Identifikation definitorischer Sätze ist. Pearsons Ziel ist nicht die Erkennung aller vorhandenen Terme. Vielmehr sollen ausschliesslich eindeutig identifizierte Begriffe der Definitionsextraktion zugrunde liegen.

4.2.1.1 Einfache formale Definitionen

Das Muster der formalen Definition nach Trimble entspricht der Gleichung

X = Y + unterscheidende Merkmale,

(29) *Telewriting is a communication technique that enables the exchange of handwritten information through telecommunication means.*[149]

wobei X dem Term, Y dem Oberbegriff und das Gleichsetzungszeichen dem verbindenden Verb entspricht. Wie bereits in Flowerdews Untersuchung waren auch bei Pearson zahlreiche Instanzen der formalen Definition in umgekehrter Form zu finden:

Y + unterscheidende Merkmale = X

(30) *One important natural activity which can lead to coastal change is known as Longshore Drift.*[150]

Zur Identifikation von einfachen formalen Definitionen legte Pearson zwei Bedingungssätze fest. Im ersten Satz werden die Platzhalter für X, Y und das Gleichsetzungszeichen spezifiziert. Der zweite Satz bestimmt die Form der Aussage insgesamt. Der erste Bedingungsteil enthält folgende Kriterien:

- X muss ein Term sein. Als Erstes muss X folglich den festgelegten Termbildungsmustern entsprechen. Wenn X auf der linken Seite der Gleichung erscheint, darf es zudem ausschliesslich mit dem unbestimmten Artikel erscheinen (im englischen Plural also ganz ohne Artikel). Diese Bedingung soll sicherstellen, dass der Term

[148] Das ITU-Korpus besteht aus einem Handbuch der International Telecommunications Union. Das GCSE-Korpus wurde von der Cobuild-Abteilung der University of Birmingham kompiliert und beinhaltet Texte aus Lehrbüchern auf der Stufe des *General Certificate of Secondary Education*. Das Nature-Korpus enhält Artikel aus dem "Nature"-Journal. (Pearson 1998, S. 64 f.)

[149] Pearson 1998, S. 145

[150] Pearson 1998, S. 136 f.

nicht als spezifische Instanz betrachtet, sondern in seiner allgemeinen Bedeutung verwendet wird. Wenn X auf der rechten Seite, also am Satzende, figuriert, kann es von einem bestimmten oder unbestimmten Artikel begleitet werden. Der bestimmte Artikel ist im Fall dieser Struktur zulässig, da er verdeutlicht, dass X eine Art von Y ist:

(31) *The period in which data is accumulated is called the test period.*[151]

- Y muss entweder ein Term sein oder zu einer vordefinierten Gruppe von generischen Begriffen (*class words*) gehören. Zu dieser Gruppe zählen beispielsweise *process* oder *method*. Y kann mit einem bestimmten, unbestimmten oder ganz ohne Artikel erscheinen. Das gilt für alle Instanzen von Y, ob auf der rechten oder der linken Seite der Gleichung.[152]
- Das X und Y verbindende Gleichsetzungszeichen kann durch verschiedene Verben oder Verbphrasen ausgefüllt werden. Die Liste der von Pearson akzeptierten Verbindungsverben (*connectives*) umfasst die Verben *be*, *be called*, *consist*, *be defined as*, *be known as*.[153] Anzumerken ist hier, dass das Verb *consist* auch eine Umfangsdefinition signalisieren kann. Pearson lässt allerdings nur Sätze zu, in denen es den Grundstoff des Begriffs spezifiziert:

  (32) *Skeletal muscle consists of bundles of muscle fibres held together by connective tissue.*[154]

Der zweite Satz von Bedingungen, der sich auf die Aussage als Ganzes bezieht, umfasst vier Spezifikationen.

- Der Gleichsetzungsteil X = Y der Definition muss einen Hauptsatz darstellen. Keine Phrasen oder Satzteile dürfen X (bzw. Y im Fall der umgekehrten Struktur) vorangestellt sein. Das soll Sätze ausschliessen, in denen eine erste einschränkende Phrase den Geltungsbereich der Definition relativiert.

  (33) *As used here, "distinct" refers to dissimilarity from other symbols compared with them visually, or aurally.*[155]

  Hier relativiert die Aussage *as used here* das Anwendungsgebiet der Definition, die damit keine Allgemeingültigkeit mehr hat.

[151] Pearson 1998, S. 138
[152] Pearson 1998, S. 138
[153] Pearson 1998, S. 141
[154] Pearson 1998, S. 147
[155] Pearson 1998, S. 141

- Das Verbindungsverb muss im Präsens Indikativ stehen. Der Gebrauch des Präsens signalisiert den allgemeinen Geltungsbereich der Definition. Modalverben werden mit Ausnahme von *can* ausgeschlossen. Ausdrücke wie X *may be defined as* schwächen die Gültigkeit der Definition. *Can* kann zwar ebenfalls eine einschränkende Wirkung haben, bezeichnet aber in Pearsons Textmaterial meist die Möglichkeit mehrerer gültiger Definitionen:

  (34) *A model can be defined as an abstraction of a reality as seen from a certain viewpoint.*[156]

  Verbindungsverben dürfen nicht negiert sein. Wie in Abschnitt 2.1.3.1 erwähnt, sind negative Definitionen wenig aussagekräftig.

- Fokusadverbien (*focusing adverbs*) unterteilt Pearson in zwei Kategorien. Die eine Gruppe gibt der Definition allgemeine Gültigkeit (*commonly, generally, usually*). Die zweite Klasse hingegen schränkt den Geltungsbereich ein (z.B. *often, mostly, primarily*, u.a.). Entsprechend akzeptiert Pearson nur die drei Adverbien der ersten Gruppe in einer Definition und schliesst Sätze wie Beispiel (35) aus:

  (35) *The cladding mode stripper often consists of a material having a refractive index equal to or greater than that of the fibre cladding.*[157]

- Als letzte Bedingung schliesslich legt Pearson fest, wie das unterscheidende Merkmal eingeleitet werden kann. Nach Y musste entweder eine Präposition (Beispiel 36), ein Relativpronomen oder ein Partizip der Vergangenheit (*past participle*, Beispiel 37) folgen.[158]

  (36) *Magnetic tape is a plastic tape with a thin coating of metal oxide […]*
  (37) *Bile is a green liquid made in your liver [...]*

Pearson hält zusammenfassend fest, dass mit Hilfe der genannten Bedingungen die meisten der einfachen formalen Definitionen identifiziert werden konnten. Der Ausschluss von Modalverben und Fokusadverbien stellt sicher, dass nur allgemein gültige Definitionen extrahiert wurden.[159] Sie stellt allerdings keine quantitativen Resultate vor.

[156] Pearson 1998, S. 142
[157] Pearson 1998, S. 142 f.
[158] Pearson 1998, S. 144 ff.
[159] Pearson 1998, S. 150

4.2.1.2 Komplexe formale Definitionen

Pearsons Diskussion komplexer definitorischer Äusserungen beschäftigt sich in erster Linie mit einigen häufig auftretenden Mustern und deren Umwandlung in einfache definitorische Sätze. Die Untersuchung beschränkt sich auf die zwei Verbindungsverben *be* und *be called*.[160] Zu den für komplexe formale Definitionen charakteristischen Mustern gehören:

- Definitorische Äusserung. *This is called […]*

  (38) *To get pure lines the plants are pollinated with their own pollen. This is called self-pollination.*

- Titel. *This is […]*

  (39) *The iris. This is a flat ring of muscle which controls the amount of light that enters the eye.*

- X. *This is a […]*

  (40) *[…] a mortgage. This is a long-term loan either at a fixed or a changing rate of interest.*[161]

Bei den letzten zwei Mustern findet sich der Begriff als Titel bzw. als das letzte Element des vorhergehenden Satzes. In diesen Fällen wird bei einer Umwandlung in eine einfache Definition das Demonstrativpronomen *this* durch X ersetzt. Im ersten Muster folgt X dem Ausdruck *this is called*. Bei einer Umformulierung wird X an den Anfang des definitorischen Satzes gestellt und durch *is a process whereby* ergänzt:

(41) *Self-pollination is a process whereby plants are pollinated with their own pollen to get pure lines.*[162]

Zusammenfassend hält Pearson fest, dass ihre Untersuchung auf nur wenigen, aber sehr produktiven Kriterien basierte. Die von ihr verwendeten Muster identifizierten eine grosse Anzahl definitorischer Äusserungen.[163]

4.2.1.3 Semiformale Definitionen

Semiformale Definitionen nach Trimble enthalten keinen Oberbegriff. Entsprechend hat die zugehörige Gleichung die folgende Form:

---

[160] Pearson 1998, S. 151
[161] Pearson 1998, S. 152 ff.
[162] Pearson 1998, S. 152
[163] Pearson 1998, S. 157

X = unterscheidendes Merkmal [164]

(42) *The output light beam has the role of the axon, broadcasting the signal from each neuron.*[165]

Die Bedingungen zur Form von semiformalen Äusserungen entsprechen zu einem grossen Teil denjenigen der formalen Definition. Die Abweichungen bestehen in einer Lockerung der Bedingungen für X und die Verbindungsverben:

- X kann unabhängig von seiner Platzierung im Satz von einem bestimmten oder unbestimmten Artikel begleitet werden. Der bestimmte Artikel wird auch für Terme am Satzanfang akzeptiert, weil bei semiformalen Definitionen der Begriff häufig bereits als Titel oder in paraphrasierter Form in Erscheinung getreten war.[166] Das kann allerdings eine korpusspezifische Beobachtung sein.
- Für Verbindungsverben in semiformalen definitorischen Sätzen fand sich in Pearsons Korpus eine grössere Anzahl von Möglichkeiten als für formale Definitionen, nämlich *contain, have, be used for, be used to, include, involve, be characterized by, be described as, produce, provide*. Pearson beschränkte sich bei ihrer Untersuchung auf die Verben *be used to, be used for, have.*[167]

In Pearsons Kopora waren semiformale Definitionen einiges zahlreicher als formale definitorische Sätze. Allerdings zeigte sich, dass der Oberbegriff häufig bereits in einer vorhergehenden Textstelle erschienen war und aus diesem Grund weggelassen wurde. Ein beträchtlicher Teil semiformaler definitorischer Sätze entsprach also eigentlich komplexen formalen Definitionen.[168]

Pearson nennt zahlreiche Ansatzpunkte zur Extraktion definitorischer Sätze. Sie spezifiziert die Art der als Begriff und Oberbegriff qualifizierenden Wörter und der sie begleitenden Artikel. Art, Modus und Tempus des Verbindungsverbs werden festgelegt. Weiter bestimmt Pearson die Form des Gesamtsatzes und schliesst vorangestellte Phrasen und Fokusadverbien aus. Schliesslich wird die einleitende Form des unterscheidenden Merk-

---

[164] Pearson 1998, S. 158
Das Gleichsetzungszeichen entspricht nicht eigentlich dem semantischen Gehalt der Verbindungsverben. Zutreffender scheint die von Meyer (2001, S. 287) gewählte Darstellung als logische Implikation: X → Merkmale

[165] Pearson 1998, S. 162

[166] Pearson 1998, S. 158

[167] Pearson 1998, S. 158

[168] Pearson 1998, S. 162

mals bestimmt. Die von Pearson verwendeten Kriterien wie Artikel- und Verbart, Einleitung des Merkmals und Ausschluss von gewissen Phrasen und Adverbien können als allgemein anwendbar betrachtet werden, auch wenn die darin enthaltenen Elemente (z.B. die Art der Verben oder Artikel) an korpus- und sprachspezifische Eigenheiten angepasst werden müssen. Basierend auf Pearsons Arbeit wurden auch bei der in Kapitel 4.4 erörterten Untersuchung deutschsprachiger Definitionskandidaten Bedingungen betreffend Artikel und Verb festgelegt. Spezifikationen zur Einleitung des Merkmals und dem Ausschluss von Phrasen und Adverbien konnten aufgrund der beschränkten Grösse des Testkorpus nicht festgelegt werden und wurden daher nicht implementiert, bieten aber interessante Ansatzpunkte.

Interessant wären genauere Angaben zur erzielten Treffergenauigkeit und zur Grösse der Trefferausbeute. Weiter könnte eine Einbeziehung von Umfangsdefinitionen die Resultate erweitern. Pearson befasst sich ausschliesslich mit der Extraktion von Inhaltsdefinitionen. Meronymie-Beziehungen, wie sie in Umfangsdefinitionen dargestellt sind, figurieren nicht in ihrer Untersuchung. Pearsons Klassifizierung definitorischer Äusserungen nach Austin scheint für eine automatische Identifikation nur beschränkt nützlich, aber die Unterscheidung zwischen formalen und semiformalen Definitionen ergibt wichtige Hinweise zur jeweiligen Satzstruktur.

Eine Pearson verwandte Arbeit, die sich allerdings stärker an der Extraktion von Begriffsbeziehungen orientiert, soll im folgenden Kapitel vorgestellt werden.

4.2.2 Meyer

Meyer befasst sich mit der Extraktion informationsreicher Kontexte (*knowledge rich contexts (KRC)*). Unter einem KRC versteht Meyer einen Kontext, der mindestens ein Element des Fachgebietswissens enthält, das für die Analyse des Begriffs relevant ist.[169] Das Element ist üblicherweise ein Merkmal oder die Darstellung einer Beziehung zwischen zwei Begriffen wie Hyponymie oder Meronymie. Basierend auf dieser Definition eines KRC entwickelt sich die Suche nach Beziehungen zwischen Begriffen, d.h. zwischen Hyperonym und Hyponym oder Teilen eines Gesamtbegriffs, zum zentralen Element in Meyers Untersuchung.

[169] Meyer 2001, S. 281

4.2.2.1 Formen informationsreicher Kontexte

Meyer unterscheidet zwischen definitorischen und erklärenden Kontexten (*defining* und *explanatory contexts*). Ein definitorischer Kontext entspricht grundsätzlich dem Muster der analytischen Definition:

X = Y + unterscheidende Merkmale

wobei X der zu definierende Begriff und Y das Hyperonym ist. Allerdings können definitorische Kontexte inhaltlich oder formal von Definitionen abweichen.[170] In inhaltlicher Hinsicht können Merkmale teilweise oder vollständig fehlen oder nicht relevant sein. Die Wahl des Oberbegriffs hängt von der Perspektive des Autors ab, und es ergibt sich nicht immer ein kohärentes Begriffssystem. Zudem findet sich häufig nicht das direkte Hyperonym, sondern ein ihm übergeordneter, allgemein gehaltener Begriff. So zeigte Meyers Korpus ein Beispiel, in dem *vermicompost* nicht als eine Art von *compost*, sondern als eine Form von *soil conditioner* definiert wurde.[171] Diese Aussagen lassen erkennen, dass sich Meyer stärker mit der terminologischen Verwendung von Definitionen beschäftigt als Pearson, die nicht auf inhaltliche Probleme eingeht. Meyer merkt an, dass formale Abweichungen von der grundsätzlichen Form der Inhaltsdefinition eine automatische Extraktion erschweren können. Der Begriff kann durch ein Pronomen oder eine generische Benennung ersetzt werden. Das Gleichsetzungselement des definitorischen Ausdrucks muss nicht lexikalisiert sein, sondern kann auch durch Satzzeichen ausgedrückt werden. So können Kommata einen Begriff und das ihm als Apposition folgende Hyperonym trennen (z.B. *compost, a dark, nutrient-rich soil conditioner, consists of […]*). Die Merkmale können adjektivisch realisiert sein und in diesem Fall vor dem Oberbegriff stehen (z.B. *nutrient-rich soil conditioner* im Sinn von *a soil-conditioner containing nutrients*).[172]

Bei erklärenden Kontexten wird kein Hyperonym angeführt. Das Y-Element der Formel entfällt, und der erklärende Kontext entspricht nicht mehr einer Gleichung. Meyer führt folgende Beispiele an:

(43) *Compost enriches topsoil with organic matter and plant nutrients, improves water infiltration, and increases water availability and nutrient retention in sandy soils.*
(44) *Compost contains nutrients, nitrogen, potassium and phosphorus.*[173]

[170] Meyer 2001, S. 283
[171] Meyer 2001, S. 285
[172] Meyer 2001, S. 286
[173] Meyer 2001, S. 287

In Beispiel (43) wird die Wirkung des Begriffs erläutert. Beispiel (44) zählt Inhaltsstoffe auf und ähnelt damit einer Umfangsdefinition.

Meyer formuliert erklärende Kontexte als logische Implikation:

$$X \rightarrow \text{Merkmale}.^{174}$$

Allerdings verwendet sie die Implikation nicht im strikten logischen Sinn von 'muss haben', sondern in der Bedeutung von 'hat typischerweise'.

Pearson und Trimble formulierten bereits einige Gründe für die Unterdrückung des Hyperonyms. Der Oberbegriff kann bereits in vorhergehenden Textstellen spezifiziert worden sein, oder der Autor setzt ihn als bekannt voraus. Meyer führt weiter an, dass die Wichtigkeit des Hyperonyms nicht in allen Fachgebieten gleich gross ist. In medizinischen Texten beispielsweise kann die Angabe der Symptome oder der Behandlungsweise wichtiger sein als die Spezifikation des Oberbegriffs. Einige Worttypen wie Verben oder Adjektive eignen sich nicht für die Form der Inhaltsdefinition, da sich die Angabe eines Hyperonyms schwierig gestaltet. Schliesslich kann der fehlende Oberbegriff auf mangelnde Schreibkompetenz des Autors zurückgeführt werden. Die unterschiedlichen Ursachen für einen fehlenden Oberbegriff zeigen, dass eine Extraktion dieser Art von definitorischen Aussagen trotz ihrer Unvollständigkeit lohnend sein kann.[175]

4.2.2.2 Methode, Muster und Einschränkungen

Meyers Verfahren setzt eine vorgängige Termerkennung voraus. Die Identifikation der ursprünglichen Extraktionsmuster basiert auf einer manuellen Analyse. Zum einen wurden Key-Word-in-Context-Konkordanzen für eine Anzahl Terme generiert und die dabei gefundenen KRC ausgewertet. Ein zweiter Ansatz bestand darin, Kontext- und Definitionsfelder einer bestehenden Terminologiedatenbank[176] zu analysieren. Dieses Vorgehen ist effizient, da die meisten Einträge relevante Muster enthalten. Allerdings ist zu beachten, dass die Kontexte sehr sorgfältig ausgewählt und damit nicht unbedingt repräsentativ sind. Zudem entfallen paralinguistische Muster, da die Feldinhalte nicht mehr in einen Text eingebettet sind.[177]

---

174 Meyer 2001, S. 287
175 Meyer 2001, S. 288
176 TERMIUM, die offizielle Datenbank der kanadischen Regierung (ca. eine Million Einträge in Englisch, Französisch und Spanisch)
177 Meyer 2001, S. 292

Meyer wendet offen gehaltene Muster an und erörtert mögliche Einschränkungen. Die Extraktionsmuster (*knowledge patterns*) können lexikalischer, syntaktischer oder paralinguistischer Natur sein. Lexikalische Muster basieren auf einem oder mehreren Schlüsselwörtern. Einige der von Meyer im Rahmen der Extraktion von definitorischen Kontexten für die Identifikation der Hyponymie-Beziehung verwendeten Muster sind *is a, classified as* und *defined as*. Die Identifikation von Meronymie basierte auf lexikalischen Elementen wie *its, is a part of* und *contains*. Diese Art von Mustern ist für erklärende Kontexte relevant.[178] Syntaktische Muster setzen ein Korpus mit Part-of-Speech-Erkennung voraus. Als Beispiel eines syntaktischen Musters zur Merkmalerkennung führt Meyer die Kombination Adjektiv + Substantiv an.[179] Zu den paralinguistischen Mustern gehören Satzzeichen und verschiedene Strukturelemente eines Texts. Strukturelemente können Formatierungen wie beispielsweise Fettdruck sein. In Meyers Korpus fanden sich auch Definitionen nach im Text formulierten Fragen.[180]

Meyer weist auf einige Problembereiche bei der Wahl der Muster hin. Ein Muster kann verschiedene Beziehungen signalisieren. Meyer führt zur Illustration das Verb *classify as* an, das sowohl ein Hyperonym (Beispiel 45) als auch ein Hyponym (46) kennzeichnen kann:

(45) *The EPA classifies composting as a form of recycling.*
(46) *Most of the compost currently being used by the landscape and nursery industries is classified as unrestricted grade.*[181]

Während einige Muster offensichtlich sind, z.B. *is a* zur Identifizierung von Hyponymie, sind andere unvorhersehbar, wie das Wort *recolonize*, das im folgenden Beispiel Mikroben als Teil des Komposts spezifiziert:

(47) *The numbers and types of mesophilic microbes that recolonize compost as it matures […]*[182]

Anzumerken ist hier, dass im vorliegenden Fall kaum von einem sich wiederholenden Element gesprochen werden kann, dass also eigentlich kein Muster vorliegt. Präzise

178 Meyer 2001, S. 290
179 Meyer 2001, S. 290
180 Meyer 2001, S. 291
181 Meyer 2001, S. 295
Meyer betrachtet *unrestricted grade compost* in Beispiel (46) als Term. Diese Einstufung ist zweifelhaft, da das Adjektiv rein beschreibend wirken kann, und eine Verbindung von Adjektiv und Substantiv nicht zwingend einen Begriff bildet.
182 Meyer 2001, S. 295

Muster zu finden kann bei einigen Begriffsbeziehungen speziell problematisch sein. Meyer verweist in diesem Zusammenhang auf die Meronymie-Beziehung. In ihrer Untersuchung zeigte sich, dass die Meronymie-Muster für *compost* schlecht auf *composting* anwendbar waren. *Composting* bezeichnet nicht ein konkretes Objekt, sondern einen Prozess, und aus diesem Grund sind seine Teile eher zeitlicher als physischer Natur. Entsprechend sind Meronymie-Muster durch lexikalische Elemente (z.B. X *is a stage of* Y) gekennzeichnet, die für *compost* nicht relevant sind.[183] Schliesslich stellt auch die Auflösung von anaphorischen Bezügen oder Umschreibungen von Begriffen ein Problem bei der automatischen Extraktion von KRC dar.[184]

Einschränkungen sind notwendig, um unerwünschte Kontexte auszuschliessen. Neben bereits von Pearson erwähnten Einschränkungen wie den Ausschluss von Modalverben weist Meyer auf die Relevanz der Grösse des Suchfensters hin, also dem Abstand zwischen Term und Muster. Ein grosses Suchfenster erhöht das Risiko unerwünschter Kontextkandidaten. Wird das Suchfenster zu klein angesetzt, schliesst das System möglicherweise relevante Kontexte aus.[185] Die Formulierung gewisser Bedingungen ist vom Ziel des Extraktionsverfahrens abhängig. Ob die folgenden Beispiele durch Ausschluss von Wörtern wie *topic* und *idea* unterdrückt werden sollen, richtet sich nach dem Anwendungsbereich des entwickelten Systems:

(48) *Composting is a traditional idea with a broad new appeal.*
(49) *Compost is a rich topic for scientific research*.[186]

Wird das System für ein bestimmtes Gebiet und eine vorgegebene Textart entwickelt, kann die beinahe vollständige Unterdrückung von unerwünschten Kandidaten realisiert werden. Soll das System hingegen generisch und auf verschiedene Fachgebiete und Textarten anwendbar sein, lassen sich ungültige Resultate kaum vermeiden. Mit der Anzahl möglicher Bedeutungen eines lexikalischen Musters steigt das Risiko, dass es unerwünschte Kandidaten liefert. Die Präposition *in* beispielsweise liefert viele Kontexte zu Meronymie-Beziehungen (z.B. *If your compost mix is too low in nitrogen […]*), aber auch zahlreiche nicht

[183] Meyer 2001, S. 298
[184] Meyer 2001, S. 297
[185] Meyer 2001, S. 291
[186] Meyer 2001, S. 293

verwendbare Textstellen (z.B. *as temperatures rise in the compost […]*). Entsprechend hoch ist in diesem Fall die Zahl der notwendigen Einschränkungen.[187]

Meyer bietet anhand einiger Beispiele Hinweise zu Problemen und möglichen Lösungen bei der automatischen Extraktion von definitorischen Sätzen. Ihre Ziele sind terminographieorientiert. Entsprechend erörtert sie im Gegensatz zu Pearson auch den Inhalt der zu identifizierenden Kontexte. Im Gegenzug fehlt eine detaillierte Diskussion der zur Suche von Extraktionskandidaten verwendeten Muster, vor allem der Muster syntaktischer Art. Einige der von ihr erwähnten Beispiele, wie die Präposition *in* als lexikalische Markierung, lassen darauf schliessen, dass die Muster grundsätzlich sehr offen angelegt sind. Da sie ihre Resultate nicht quantifiziert, gestaltet sich die Beurteilung des Erfolgs ihres Vorgehens schwierig. Ihr Extraktionsverfahren richtet sich an der Identifikation von Begriffsbeziehungen aus, und sie erörtert nicht nur Hyponymie, sondern auch Meronymie. Die Untersuchung von Umfangsdefinitionen, denen Meronymiemuster zugrunde liegen, und Meyers Ansatz, zur Verbesserung der Suchresultate gewisse korpusspezifische Wörter auszuschliessen, wurden in der in Kapitel 4.4 erörterten Untersuchung definitorischer Sätze im Deutschen übernommen. Interessant ist Meyers Berücksichtigung paralinguistischer Muster. Diese Einschränkung konnte allerdings aufgrund mangelnder Formatierung in der Studie zu deutschsprachigen Definitionen nicht berücksichtigt werden.
Im folgenden Abschnitt wird eine Untersuchung vorgestellt, die Meyers Erkenntnisse verwendet, aber stärkeres Gewicht auf die Einschränkung von Mustern legt.

4.2.3 Rebeyrolle

Rebeyrolle konzentriert sich in ihrer Untersuchung auf ein Muster zur Identifikation von Inhaltsdefinitionen und dabei vor allem auf die Ausarbeitung von Einschränkungen. Ihre Arbeit beinhaltet eine detaillierte Erörterung der für ein französisches Testkorpus entwickelten Restriktionen und deren Anwendung.

Rebeyrolle unterscheidet zwischen lexikalisch-syntaktischen, typographischen und satzübergreifenden Musterbeschränkungen. Ihr Ziel ist die Identifikation definitorischer Äusserungen der folgenden Form:

[187] Meyer 2001, S. 292

N_a *être* N_x – X,

wobei N_a dem Term, N_x dem Oberbegriff und X dem einschränkenden Merkmal entspricht:

(50) *Un carré est une figure à quatre côtés égaux et à angles droits.*[188]

Rebeyrolle führt keine Termerkennung durch, und N_a und N_x unterliegen daher nur der Bedingung, Substantive zu sein. Ihre lexikalisch-syntaktischen Einschränkungen stützen sich auf die Untersuchung von Pearson. Die Art des mit Term und Oberbegriff erscheinenden Artikels wird bestimmt, ebenso die Form des Prädikats und die Einleitung der einschränkenden Merkmale. Interessant ist ihre Beobachtung, dass das Hyperonym N_x von klassifizierenden Ausdrücken wie *sorte, espèce, variété, type* begleitet werden kann, die semantisch ambig sind. Diese Elemente können einerseits eine hierarchische Ober- / Unterbegriff-Beziehung verdeutlichen, andererseits eine nur approximative Ähnlichkeit ausdrücken und damit die Gültigkeit der Aussage abschwächen.[189] Da Rebeyrolle keine Termerkennung durchführt, liefert ihre Grundstruktur zahlreiche Instanzen, in denen N_x nicht einem Oberbegriff entspricht (z.B. N_a *est un exemple*). Mit Hilfe von Stopplisten können solche Vorkommen ausgeschlossen werden. Das erfordert allerdings eine sorgfältige und korpusspezifische Auswahl der zu unterdrückenden Substantive.[190]

Zur zweiten Kategorie, den Einschränkungen drucktechnischer Art, hält Rebeyrolle fest, dass in dem von ihr untersuchten Korpus Terme häufig typographisch gekennzeichnet waren (z.B. Fettdruck oder Grossbuchstaben).[191] Im Rahmen der dritten Kategorie, den satzübergreifenden Beschränkungen, untersucht Rebeyrolle die von Pearson und Flowerdew gemachte Feststellung, dass das Definiendum in vorhergehenden Textstellen bereits erwähnt worden sein kann. Rebeyrolle befasst sich nicht mit pronominalen Wiederaufnahmen, sondern mit Wiederholungen des Terms selbst, die gleichzeitig als ein Indiz für dessen Relevanz gewertet werden. Zwei häufige Formen von Wiederholungen liessen sich in ihrem Korpus feststellen. Der Term erscheint entweder innerhalb des vorstehenden Titels oder im vorhergehenden Satz. Die Repetition erfolgt meist innerhalb zweier aufeinander folgender Sätze; grössere Distanzen sind selten.[192]

---

[188] Rebeyrolle 2000, S. 105 f.
[189] Rebeyrolle 2000, S. 107
[190] Rebeyrolle 2000, S. 112
[191] Rebeyrolle 2000, S. 108
[192] Rebeyrolle 2000, S. 110

Zur Erkennung gültiger Satzkandidaten wendet Rebeyrolle ein hierarchisch aufgebautes Filtersystem an. In einem ersten Schritt wird nach einem Grundmuster, in diesem Fall dem Verb *être* in der dritten Person Indikativ gefolgt von einem unbestimmten Artikel, gesucht. Die Einschränkungen werden in folgender Reihenfolge angewendet:

1. lexikalisch-syntaktische Beschränkungen

   Als erster Filter dient die Spezifikation einer geforderten lexikalisch-syntaktisch Struktur, die im vorliegenden Fall die folgende Form aufweist:

   {*le, la, les, l', un*} N_a {*est, sont*} {*un, des*} N_x

2. a) satzübergreifende Bedingungen

   Findet sich im vorhergehenden Satz oder Titel bereits ein Vorkommen von N_a, wird der Satz als gültig eingestuft und das Vorgehen abgebrochen.

2. b) typographische Bedingungen

   Erscheint N_a nicht im vorstehenden Text, finden typographische Bedingungen ihre Anwendung.[193]

Rebeyrolle erörtert nur ein Muster zur Extraktion definitorischer Kontexte. Entsprechend sind die von ihr spezifizierten Einschränkungen auf eine Form der Inhaltsdefinition zugeschnitten. Sie quantifiziert die Resultate, so dass eine Beurteilung der Effektivität ihrer Einschränkungen möglich ist. Die von ihr verwendeten lexikalisch-syntaktischen und typographischen Kriterien finden sich in ähnlicher Art bei Pearson und Meyer; die Berücksichtigung satzübergreifender Kriterien ist jedoch neu. Grundlegende Unterschiede ihres Verfahrens gegenüber dem Vorgehen von Pearson und Meyer bestehen darin, dass Rebeyrolle keine vorgängige Termerkennung anwendet und die Einschränkungen stufenweise einführt. Die graduelle Verwendung von Beschränkungen scheint erfolgversprechend. Aufgrund der beschränkten Anzahl implementierter Bedingungen konnte sie in der in Kapitel 4.4 diskutierten Extraktion deutscher Definitionskandidaten nicht berücksichtigt werden. Übernommen wurde hingegen der Verzicht auf eine Termidentifikation.

Pearson, Meyer und Rebeyrolle konzentrieren sich auf die Ausarbeitung von Mustern und Restriktionen zur Identifikation von Definitionen. Das ist auch ein Anliegen der zwei

[193] Rebeyrolle 2000, S. 112

weiteren nachfolgend diskutierten Arbeiten. Sie befassen sich vor allem aber auch mit der Darstellung der semantischen Komponenten von Definitionen.

4.2.4 Bowden et al.

Bowden et al. unterscheiden drei Formen von Begriffserläuterungen, die Definition, die Exemplifikation (*exemplification*) und die Partition (*partition*). Bei der Exemplifikation wird ein Begriff mit Hilfe eines oder mehrerer Beispiele veranschaulicht. Die Partition erläutert einen Begriff durch eine (nicht notwendigerweise vollständige) Liste seiner Bestandteile und ist damit der Umfangsdefinition ähnlich. Behandelt werden Begriffserläuterungen innerhalb eines Satzes.[194] Das Vorgehen von Bowden et al. umfasst drei Schritte: eine Triggering-Stufe, innerhalb derer definitorische Sätze anhand von Schlüsselwörtern identifiziert werden, die Extraktion der Begriffsbeziehungen durch Vergleiche mit nicht syntaktischen Mustern und eine Validierung der extrahierten Sätze mit Hilfe einer syntaktischen Analyse.

Im Triggering-Schritt werden Sätze aufgrund darin enthaltener Schlüsselwörter oder -ausdrücke als Kandidaten gekennzeichnet. Er basiert also auf lexikalischen Elementen. Bei Definitionssätzen sind zwei mögliche Auslöser die Formen *define* und *is a*. Gleichzeitig existieren negative Trigger, die den Satz als nicht valid markieren. Bei der Definition ist das beispielsweise der Ausdruck *there is a*. Pro Erläuterungsform existieren ein positives und ein negatives Trigger-Set mit ca. je 20 Einträgen.[195] Nach dieser ersten Phase werden die Satzkandidaten in einem nicht syntaktischen Mustervergleich behandelt. Darin werden die Satzglieder mit Token versehen, die ihre semantische Funktion innerhalb der Erläuterung bezeichnen. So steht das Token **e** für *an example of*, = für das Verb *be* und **X** für Wörter, denen keine Funktion zugeordnet wird. Satzzeichen werden als solche dargestellt. Alle möglichen Segmentierungen werden generiert. Bowden et al. veranschaulichen dieses Vorgehen am folgenden Beispiel:

(51) *An example of a high-level language is PASCAL.*[196]

[194] Bowden et al. 1996, S. 150
[195] Bowden et al. 1996, S. 151
[196] Bowden et al. 1996, S. 152

Die möglichen Segmentierungen haben folgende Form:

1) eX=X. eX=X
2) eX. eX
3) X=X. X=X
4) X. X

In Fall (1) bilden die Token **e** und **=** Fixpunkte. Entsprechend wird Beispielsatz (51) wie folgt aufgelöst:

e → A*n example of*
X → *a high-level language*
= → *is*
X → *PASCAL(.)*

Fall (2) stützt sich auf das Token **e** und summiert den restlichen Satz unter **X**. Daraus ergibt sich die folgende Segmentierung:

e → A*n example of*
X → *a high-level language is PASCAL(.)*

Fall (3) betrachtet das Gleichsetzungszeichen als entscheidendes Element und zeigt folgende Struktur:

X → *An example of a high-level language*
= → *is*
X → *PASCAL(.)*

In Fall (4) schliesslich wird der gesamte Satz *An example of a high-level language is PASCAL(.)* unter **X** zusammengefasst.

Wie aus der oben stehenden Segmentierung ersichtlich wird, kann das Token **X** auch für anderen Token zugeordnete Wörter stehen, wenn diese Funktionstoken in der aktuellen Segmentierung nicht verwendet werden. Die so erhaltenen Segmentierungsvorschläge werden mit einer Vorlage verglichen, die für die jeweiligen Sätze zulässige semantische Muster enthält. In Beispiel (51) findet sich eine Übereinstimmung mit einem für die Exemplifikation spezifizierten Muster **eC=0**. Auch hier bezeichnet **e** die Beziehung *example*. **C** steht für *concept*. Die extrahierten Beispiele werden mit Ordnungszahlen versehen; aus diesem Grund ist das vorliegende Beispiel *PASCAL* mit 0 gekennzeichnet. Der Satz entspricht also folgendem Muster:

e → *an example of*
C → *a high-level language*
= → *is*
0 → *PASCAL*[197]

[197] Bowden et al. 1996, S. 153

Im dritten Schritt werden die Textfragmente der vorherigen Phase nach einer vorgängigen Part-of-Speech-Erkennung syntaktisch analysiert. Auch dieses Vorgehen basiert auf Mustervergleichen.[198] Die Struktur des Satzkandidaten wird mit syntaktischen Mustern von Begriffserläuterungen verglichen. Bowden et al. erörtern keine Beispiele, halten jedoch fest, dass das Muster-Set weiter ausgearbeitet werden soll.[199]

Der Vorteil des Systems von Bowden et al. liegt in seiner Fachgebietsunabhängigkeit. Sie halten fest, dass in den spezifizierten Mustern keinerlei sachgebietsabhängige Information kodiert ist. Der zweite Schritt in ihrem Verfahren, die Unterteilung in Tokens und die Generierung aller möglichen Lesarten, ist jedoch aufwändig. Allerdings wird die Weiterverwendung der so erhaltenen Informationen durch die dadurch erzeugte strukturierte Form erleichtert. Der von Bowden et al. gewählte Ansatz unterscheidet sich von den in den Abschnitten 4.2.1 bis 4.2.3 vorgestellten Methoden dadurch, dass primär eine semantische Segmentierung vorgenommen wird und syntaktische Informationen erst in einer letzten Phase einbezogen werden. Die semantischen Komponenten von definitorischen Sätzen werden explizit dargestellt. Eine mögliche Verwendung ihres Verfahrens sehen Bowden et al. in der Wissensakquisition für semantische Netze. Da der Fokus sowohl von Bowden et al. als auch von der im folgenden Abschnitt erörterten Arbeit von Büchel et al. auf der strukturierten Darstellung der Komponenten einer Definition liegt und nicht auf der Definitionsextraktion als solche, wurden keine Elemente dieser beiden Studien in der Arbeit zu deutschsprachigen Definitionskandidaten berücksichtigt.

4.2.5 Büchel et al.

Büchel et al. beschreiben ein Verfahren zur Bedeutungsbestimmung von Begriffen in deutschen Fachtexten anhand syntaktischer Muster. Ihr Ziel ist es, definitorische Sätze zu extrahieren, sie in Definiendum, Definiens und Kopula zu segmentieren und aus diesen Komponenten ein semantisches Netz zu generieren. Der Schwerpunkt liegt nicht so sehr auf der Extraktion aller möglichen definitorischen Sätze, sondern auf der strukturierten Darstellung der Begriffsbestimmung. Sie implementieren ihr Verfahren in einem ersten Schritt denn auch unter Berücksichtigung nur eines syntaktischen Musters.

---

[198] Bowden et al. 1996, S. 153
[199] Bowden et al. 1996, S. 154

Büchel et al. basieren ihre Muster auf einer Untersuchung von Weber.[200] Er analysiert und klassifiziert morphologische und syntaktische Strukturen von Definitionstexten in einem einsprachigen deutschen Wörterbuch[201]. Ein Grossteil der Muster eignet sich daher nicht für die Extraktion kontextualisierter Definitionen. Das in der vorliegenden Untersuchung verwendete Muster besteht aus einer Nominalphrase mit nachfolgendem Relativsatz.[202] Dieses Muster wird nach einer Termerkennung auf ein maschinenlesbares Korpus[203] angewendet. Begriffe werden im Korpus auf der Basis einerseits eines Schlagwortverzeichnisses, andererseits drucktechnisch betonter Wortformen identifiziert.[204] In einem ersten Schritt extrahieren Büchel et al. Sätze, in denen ein zu untersuchender Begriff belegt ist. Durch automatisches Tagging wird jede Wortform der Begriffsumgebung mit einer Wortklassenangabe versehen.[205] Schliesslich identifiziert ein maschineller Abgleich mit dem Wortklassenmuster für Begriffsbestimmungen bedeutungsrelevante Stellen, die in Definiendum, Definiens und Kopula segmentiert werden.[206] Da Büchel et al. sich wie oben erwähnt primär mit der Generierung von semantischen Netzen befassen, finden sich keine Angaben zum Erfolg des von ihnen verwendeten Musters oder zur Anwendbarkeit anderer von Weber identifizierten syntaktischen Strukturen.

4.3 Identifikation von Begriffsbeziehungen

Um den Überblick über die Arbeiten zur Extraktion definitorischer Sätze durch Einbezug eines verwandten Gebiets zu vervollständigen, folgt nachstehend eine kurze Erörterung von Untersuchungen zur Identifikation von Begriffsbeziehungen.

Zahlreiche Arbeiten befassen sich mit lexikalisch-syntaktischen Mustern, die Beziehungen zwischen Begriffen kennzeichnen. Häufig untersuchte Beziehungen sind Meronymie und Hyponymie. Im Rahmen des WordNet-Projekts[207] befasst sich Hearst[208] mit der Identifi-

---

200 Weber 1993
201 Duden Deutsches Universalwörterbuch 1989
202 Büchel et al. 1995, S. 132
203 Hegels Text "Wissenschaft der Logik"
204 Büchel 1996, S. 276
205 Büchel et al. 1995, S. 135. Als Tagging-Verfahren wurde LEMMA2_C verwendet, eine Portierung des Algorithmus LEMMA2 in die Programmiersprache C unter dem Betriebssystem UNIX. Büchel et al. machen keine Angaben zur Art des Verfahrens (z.B. zur Berücksichtigung von Mehrdeutigkeiten).
206 Büchel et al. 1995, S. 135
207 WordNet ist eine lexikalische Datenbank.
208 Hearst 1992

kation von Ober- und Unterbegriffen, und Miller[209] erörtert die in WordNet kodierten Meronymie-Muster. Borillo[210] und Otman[211] untersuchen die automatische Erkennung von Hyponymie bzw. Meronymie im Französischen. Das von Hearst entwickelte Verfahren zur Identifikation von relevanten Mustern basiert auf der Analyse von Textstellen, die Paare von Ober- und Unterbegriffen enthalten. Die darin feststellbaren Gesetzmässigkeiten werden für die Identifikation neuer Begriffspaare verwendet. Hearsts Vorgehen wurde von Morin[212] und Rousselot[213] in den Systemen *Prométhée* und *Reltex* automatisiert. In Morins System *Prométhée* werden Begriffspaare eingespeist, zwischen denen die gewünschte semantische Beziehung besteht. Die Strukturen der auf dieser Basis extrahierten Sätze werden durch ein automatisches Verfahren auf einen gemeinsamen Nenner reduziert. Ein Experte validiert die so erhaltenen Muster. Die Muster sind offen gehalten und entsprechend sehr produktiv; der sich daraus ergebende Nachteil ist eine grosse Anzahl ungültiger Kandidaten.[214]

Andere Systeme weiten den Kreis der untersuchten Beziehungen aus. Ahmad et al. untersuchen neben Hyponymie und Meronymie auch Synonymie, Ursache/Wirkung und Material.[215] Jouis befasst sich ebenfalls mit verschiedenen Beziehungstypen.[216] Um die Zuverlässigkeit der extrahierten Sätze sicherzustellen, sind die in seinem System SEEK verwendeten Muster stark einschränkend und teilweise wenig produktiv.[214] Auf ein Fachgebiet zugeschnittene Muster wurden von Riloff entwickelt, die sich mit der Extraktion von Beziehungen ohne vorgängige Wortklassenannotierung befasst.[217] Sie identifiziert in einem Korpus zum Thema Terrorismus die durch Verben gekennzeichneten Beziehungen zwischen Nominalphrasen.

---

209 Miller 1999
210 Borillo 1996, zitiert nach: Aussenac 2000, S. 7
211 Otman 1996, zitiert nach: Aussenac 2000, S. 7
212 Morin 1999
213 Rousselot 1996
214 Morin und Jouis untersuchen in späteren Arbeiten die Verbindung ihrer Verfahren mit anderen Systemen. Im Fall von Morin handelt es sich dabei um eine Kopplung mit EAGLE, einem induktiven Lernsystem. Jouis untersucht die Anbindung an ein statistikbasiertes System (CONTERM).
215 Ahmad et al. 1992, zitiert nach: Meyer 2001, S. 290
216 Jouis 1993, zitiert nach: Le Priol 2000, Teil 2, Kap. 4
217 Riloff 1996

Im Gegensatz zu Riloff oder Jouis basieren Aussenac et al.[218] und Condamines et al.[219] ihre Untersuchungen auf allgemeinen Mustern, die aber durch Präzisierung auf spezialisierte Korpora zugeschnitten werden. Neben der Anpassung offener Muster ist ihr Ziel auch die Identifikation von sachgebietsrelevanten Beziehungen. Generische Muster für Hyponymie und Meronymie werden auf ein Korpus projiziert und ihre Wirksamkeit evaluiert. Das ermöglicht die Entwicklung von korpusspezifischen Einschränkungen. Das System *Caméléon* von Aussenac et al. integriert zudem die gefundenen Beziehungen in ein Sachgebietsmodell. Die Resultate von Condamines et al. werden für den Aufbau einer Wissensdatenbank verwendet.

Die Vorgehen zur Extraktion von Begriffsbeziehungen und von Definitionskandidaten und in beschränktem Mass die bei beiden Verfahren verwendeten Muster weisen Ähnlichkeiten auf. Da Hyponymie und Meronymie die Grundlage der Inhalts- und Umfangsdefinition sind, liefert die Identifikation von Ober- und Unterbegriffen oder Teil-Ganzes-Beziehungen auf jeden Fall ergänzende Angaben zur Formulierung von Definitionen.
Im folgenden Kapitel 4.4 sollen auf der Basis der in den vorhergehenden Abschnitten erörterten Literatur einige Muster zur Extraktion definitorischer Sätze aus deutschsprachigen Texten diskutiert werden.

[218] Aussenac et al. 2000
[219] Condamines et al. 2001

4.4 Muster zur Extraktion definitorischer Sätze im Deutschen

Das vorliegende Kapitel enthält eine Darstellung selbst entwickelter Muster und Restriktionen zur Erkennung von Definitionskandidaten in deutschen Texten. Die genauen Ziele sind in Abschnitt 4.4.1 dargestellt. Das Projekt basiert auf einer vorgängigen Textbearbeitung durch Part-of-Speech-Tagging, Lemmatisierung und Kennzeichnung von Nominal- und Präpositionalphrasen. Die Textvorbereitung ist in Abschnitt 4.4.2 beschrieben. Abschnitt 4.4.3 diskutiert die manuelle Bestimmung typischer Formen von Inhalts- und Umfangsdefinitionen. In den Teilen 4.4.4 und 4.4.5 werden die implementierten Muster und Beschränkungen und die damit erzielten Resultate erörtert. Anhand einer zweiten manuellen Analyse fanden sich zusätzliche Definitionsformen (Abschnitt 4.4.6), die Bestandteil weiterführender Arbeiten sein könnten (4.4.7).

4.4.1 Ziel

Ziel des vorliegenden Verfahrens war es, Muster für die Extraktion definitorischer Sätze aus dem Deutschen anhand eines technischen Handbuchs zu identifizieren. Das Projekt versteht sich als Ergänzung zur in Kapitel 3 beschriebenen Untersuchung der Siemens-Terminologiedatenbank. Es sollte die Möglichkeit geprüft werden, fehlende definitorische Elemente durch ein halbautomatisches Verfahren zu ergänzen. Zu diesem Zweck wurden auf der Basis einer manuellen Evaluation einige Muster definiert und anhand eines Testkorpus evaluiert. Das Verfahren beinhaltet keine vorgängige Termerkennung. Dies ermöglichte die gleichzeitige Identifikation von zusätzlichen Begriffen und damit die Ergänzung von Begriffslücken in der Datenbank.

Zwei grundsätzliche Muster, jeweils der Inhalts- und der Umfangsdefinition entsprechend, wurden untersucht. Ziel der Identifikation von Inhaltsdefinitionen war es, Hyperonym und Merkmale eines Begriffs zu finden. Eine inhaltliche Beurteilung der so gefundenen Kandidaten nach terminologischen Kriterien fand nicht statt. Es wurde also nicht überprüft, ob die Unterscheidungsangabe vollständig war, und nicht verifiziert, ob das Hyperonym der nächstfolgenden Stufe entstammte. Eine solche Evaluation hätte die Ausarbeitung eines in sich geschlossenen Begriffssystems vorausgesetzt. Bei den Mustern zur Identifikation von einer Umfangsdefinition ähnlichen Aussagen wurde nach Sätzen gesucht, die Teile eines

Begriffs auflisteten. Auch hier wurde nicht geprüft, ob sämtliche Komponenten eines Begriffs identifiziert waren.

4.4.2 Textvorbereitung

4.4.2.1 Testkorpus

Das Testkorpus setzt sich zusammen aus fünf Kapiteln des von der Firma Siemens auf CD-ROM zur Verfügung gestellten Handbuchs zum PCS7-System (*Process Control System*). Das Handbuch wird für Instruktionszwecke zur Verfügung gestellt und ist entsprechend sorgfältig formuliert. Es ist von Experten geschrieben und richtet sich an Fachleute des Informatikbereichs, die mit dem PCS7-System nicht vertraut sind. Die darin enthaltenen Texte weisen einen hohen Grad an Informationsdichte auf, da sie zur Wissensvermittlung angelegt sind. Neben der Beschreibung des Systems enthalten sie zahlreiche Programmbeispiele zu Musterfällen. Eine starke Textstrukturierung mit zahlreichen Betitelungen ist zu beobachten, wie sie oft in Handbüchern anzutreffen ist.
Die Kapitel standen als Dateien im PDF-Format zur Verfügung und wurden im Rahmen eines weiteren Projekts für die Firma Siemens[220] in Textformat konvertiert. Diese Umwandlung ermöglichte die Bearbeitung der Dateien wie nachfolgend beschrieben, bedingte aber auch den Verlust von typographischen Informationen wie Fettschrift oder tabellarische Darstellungen. Die fünf Kapitel umfassen insgesamt rund 170'000 Wörter.

4.4.2.2 PoS-Erkennung

Beim Testkorpus wurde eine Part-of-Speech-Erkennung (PoS-Tagging) durchgeführt, d.h. die Wortformen des Textes wurden automatisch mit Wortklassen annotiert. Die Wortklassenauszeichnung ermöglicht die Suche nach syntaktischen Mustern und bildet die Basis der in Abschnitt 4.4.2.4 diskutierten Identifikation von Nominal- und Präpositionalphrasen. Die PoS-Erkennung wurde unter Verwendung des Tree-Tagger[221] durchgeführt, ergänzt durch ein Modul zur Kennzeichnung von Eigennamen[222]. Dieses Modul erkennt und klassifiziert Eigennamen von Personen, geographischen Orten, Firmen und Produkten, die bei der automatischen PoS-Auszeichnung zu Problemen führen können. Der Tree-Tagger basiert auf dem STTS (Stuttgart-Tübingen Tag-Set), einem Tagset für die deutsche

[220] Unter der Leitung von Martin Volk, Institut für Computerlinguistik der Universität Zürich
[221] Schmid et al. 1996, zitiert nach: Volk 2001, S. 54 ff.
[222] Volk 2001, S. 54 ff.

Sprache mit rund 50 PoS-Kennzeichnungen und drei Tags für Satzzeichen.[223] Das STTS unterscheidet zwischen Eigennamen und regulären Substantiven, Voll-, Modal- und Hilfsverben und zwischen Präpositionen, kontrahierten Präpositionalformen und Postpositionen. Der Tagger bearbeitet vertikalen Text, d.h. jedes Wort und jedes Satzzeichen erscheint in einer einzelnen Zeile. Er annotiert jedes Wort innerhalb eines Satzes mit einer Kennzeichnung, ändert aber keinen vorher zugeteilten Tag. Dadurch bleibt die vorhergehende Auszeichnung für Eigennamen erhalten.[224]

4.4.2.3 Lemmatisierung

In einem zweiten Schritt wird das Testkorpus lemmatisiert. Die Substantive werden auf ihre Form im Nominativ Singular reduziert, die Verben auf ihre Infinitivform. Adjektive sind in ihrer unflektierten Stammform aufgeführt. Die Rückführung der Wörter auf ihre Stammformen macht die Implementierung lexikalischer Muster möglich, die bei der Suche nach Inhalts- und Umfangsdefinitionen angewendet wurden und in den Abschnitten 4.4.4.2 und 4.4.4.3 erörtert sind. Zur Lemmatisierung wurde das Morphologieanalysesystem GERTWOL benutzt.[225] GERTWOL ist ein rein wortbasiertes Analysesystem, das jede mögliche Lesart einer Wortform auflistet. So wird beispielsweise bei der Wortform *Junge* eine adjektivische (wie in *junge Frauen*) und eine substantivische Interpretation (*Junge* = 'männliches Kind') aufgeführt. Daher ist es nötig, das GERWTOL-Resultat mit der PoS-Kennzeichnung zu vergleichen, um das korrekte Lemma zu finden.[226]

4.4.2.4 Erkennung von Nominal- und Präpositionalphrasen

Zur Erkennung von Nominalphrasen (NP) und Präpositionalphrasen (PP) wird ein Mustervergleichssystem verwendet, das auf PoS-Mustern basiert.[227] Die Auszeichnung von NP und PP bildet die Basis der Suche nach syntaktischen Mustern, wie sie für die Identifikation von Inhaltsdefinitionen verwendet wurden und in Abschnitt 4.4.4.2 beschrieben sind. Die Phraseninformation wird im NEGRA-Exportformat gespeichert.[228] NEGRA ist ein zeilenbasiertes Format, das numerische Erkennungen für verschachtelte Phrasen verwendet. Das NEGRA-Format hat zum Ziel, die Strukturen so flach wie möglich zu halten,

[223] Thielen et al. 1996
[224] Volk 2001, S. 65
[225] Lingsoft-Oy 1994, zitiert nach: Volk 2001, S. 65
[226] Volk 2001, S. 65 f.
[227] Volk 2001, S. 69
[228] Skut et al. 1997, zitiert nach: Volk 2001, S. 69

ohne Informationen zu verlieren.[229] Im Folgenden ist ein Beispiel eines Satzes im NEGRA-Format aufgeführt.

```
Allgemeine                ADJA  -- NK   501   %% allgemein
Kenntnisse                NN    -- NK   501   %% Kennt~nis
auf                       APPR  -- AC   500   %% auf <Loc1>
dem                       ART   -- NK   500   %% <Loc2>
Gebiet                    NN    -- NK   500   %% Gebiet <Loc3>
der                       ART   -- NK   502
Automatisierungstechnik   NN    -- NK   502   %% Automat~is~ier~ung\s#techn~ik
sind                      VAFIN -- --   0     %% sein
dabei                     PROAV -- --   0
hilfreich                 ADJD  -- --   0     %% hilf|reich
.                         $.    -- --   0
#500                      PP    -- --   0
#501                      NP    -- --   0
#502                      NP    -- --   0
```

Die Information im NEGRA-Format ist in zwei Blöcke geteilt. Der erste Block enthält die Wörter und die dazugehörigen Informationen, der zweite Block die Phrasenknoten. Innerhalb des ersten Blocks erscheinen in der ersten Kolonne die Wortformen und Satzzeichen. In der zweiten Kolonne werden die PoS-Kennzeichnungen aufgeführt. Die dritte Kolonne ist reserviert für morphologische Informationen, die in diesem Fall nicht verwendet wurden. Die vierte Kolonne zeigt die Funktion des Worts innerhalb des nächsthöheren Knotens. Die fünfte Kolonne listet die nummerischen Kennzeichnungen auf, die auf die Informationen im zweiten Block verweisen. Die letzte Kolonne kann einen Kommentar (durch "%%" gekennzeichnet) enthalten. Im vorliegenden Fall wurde diese Kolonne für das Lemma verwendet. Sämtliche Konstituenten sind im zweiten Block aufgeführt. Im obigen Beispiel erkennt der Parser zwei NP und eine PP. Der Parser versucht keine Anbindungen, d.h. weder Nominalphrasen im Genitiv noch Präpositionalphrasen werden anderen Phrasen zugeordnet.

In einem letzten Schritt werden die Satzteile mit Hilfe eines von Volk entwickelten Clause-Boundary-Detector abgegrenzt.[230]

[229] Volk 2001, S. 69
[230] Volk 2001, S. 73 ff.

4.4.3 Manuelle Bestimmung von Inhalts- und Umfangsdefinitionskandidaten

Zur Projektvorbereitung wurde als Erstes ein Handbuchkapitel (43'500 Wörter) manuell auf definitorische Sätze in der Form einer Inhalts- oder Umfangsdefinition überprüft. Da sich das Projekt auf einfache definitorische Sätze, also Definitionen innerhalb eines Satzes, beschränkt, betrachtete die manuelle Bestimmung nur diese Instanzen. Auch fand keine genaue inhaltliche Prüfung der Definitionen statt. Insgesamt wurden 56 Sätze als einer Inhalts- oder einer Umfangsdefinition entsprechend eingestuft. Nachstehend folgt die Beschreibung der gefundenen Strukturen.

4.4.3.1 Inhaltsdefinition

Es fanden sich 42 Instanzen von Inhaltsdefinitionen mit fünf grundsätzlichen syntaktischen Strukturen:

- NP_{Term} + *sein* + $NP_{Hyperonym}$ + Relativsatz

  Mit 27 Vorkommen wird das Merkmal am häufigsten durch einen Relativsatz ausgedrückt.

  (52) *Lokale Variablen und Bausteinparameter sind Daten,* ***die innerhalb eines Codebausteins vereinbart werden und nur für diesen Codebaustein Gültigkeit haben****.*

- NP_{Term} + *sein* + $NP_{Hyperonym}$ mit nachgestellter PP

  In drei Sätzen wird das Merkmal in Form einer dem Oberbegriff nachgestellten Präpositionalphrase angefügt.

  (53) *Eine SCL-Anweisung ist eine ausführbare Aktion* ***im Anweisungsteil eines Codebausteins****.*

- NP_{Term} + *sein* + $NP_{Hyperonym}$ mit durch vorangestellte PP ergänztem partizipialem Adjektiv

  In weiteren drei Fällen wird das Merkmal durch eine partizipiale Adjektivphrase ausgedrückt, die durch eine vorangestellte Präpositionalphrase ergänzt wird. Wie das nachstehende Beispiel zeigt, können weitere Informationen in einer zusätzlichen Präpositionalphrase folgen.

  (54) *Ein Vergleichsausdruck ist ein* ***mit Vergleichsoperatoren gebildeter*** *Ausdruck vom Typ BOOL.*

- NP_{Term} + *sein* + $NP_{Hyperonym}$ mit Genitiv-NP

  Zwei Sätze enthalten das Merkmal in Form einer dem Oberbegriff folgenden Nominalphrase im Genitiv.

  (55) *Bausteinparameter sind formale Parameter **eines Funktionsbausteins oder einer Funktion**.*

- NP_{Term} + *sein* + AP + $NP_{Hyperonym}$

  In einem Fall wurde eine Adjektivphrase als Merkmal eingestuft.

  (56) *Das OK-Flag ist eine **systemvereinbarte** Variable.*

Als häufigstes Verbindungsverb fand sich *sein*. Sechs Sätze basierten auf den Verben *bezeichnen* und *nennen.* Diese Verben werden entweder im Passiv oder zusammen mit dem unpersönlichen *man* verwendet. Das Merkmal hat ebenfalls die Form eines Relativsatzes oder einer Präpositionalphrase, aber die syntaktische Struktur des Gesamtsatzes wird umgestellt:

- $NP_{Hyperonym}$ + Relativsatz + *bezeichnet man als/nennt man* + PP_{Term}/NP_{Term}

  $NP_{Hyperonym}$ mit PP + *bezeichnet man als/nennt man* + PP_{Term}/NP_{Term}

  Wird das Verb mit dem unpersönlichen *man* verwendet, erscheinen der Oberbegriff und das Merkmal am Satzanfang. Das Merkmal kann durch einen Relativsatz (Beispiel 57) oder eine Präpositionalphrase (Beispiel 58) realisiert werden.

  (57) *Alle Parameter, die im Vereinbarungsblock VAR_INPUT, VAR_OUTPUT und VAR_IN_OUT eines Codebausteins aufgelistet sind, bezeichnet man als Formalparameter.*

  (58) *Die Platzhalter im Baustein bezeichnet man als Formalparameter.*

- $NP_{Hyperonym}$ + Relativsatz + *wird (als)* + PP_{Term}/NP_{Term} + *bezeichnet/genannt*

  $NP_{Hyperonym}$ mit PP + *wird (als)* + PP_{Term}/NP_{Term} + *bezeichnet/genannt*

  Bei Passivkonstruktionen wird der Satz ebenfalls durch den Oberbegriff und das Merkmal in Form eines Relativsatzes (Beispiel 59) oder einer Präpositionalphrase (Beispiel 60) eingeleitet.

  (59) *Ein Bezeichner, dessen Wert während der Programmdurchführung geändert werden kann, wird Variable genannt.*

  (60) *Die Parameter in den Aufrufen innerhalb des Anweisungsteils werden Aktualparameter genannt.*

4.4.3.2 Umfangsdefinition

Die Umfangsdefinition war seltener als die Inhaltsdefinition. Insgesamt wurden 14 Vorkommen identifiziert, bei denen sich verschiedene Verben unterscheiden lassen:

- *bestehen aus* (8 Vorkommen)

  (61) *Ausdrücke **bestehen aus** Operatoren und Operanden.*

- *enthalten* (3 Vorkommen)

  (62) *Der Datenbaustein DB **enthält** globale anwenderspezifische Daten.*

- *sich zusammensetzen aus* (1 Vorkommen)

  (63) *Der Absolutbezeichner **setzt sich aus** dem Operandenkennzeichen mit Speicher- und Grössen-Präfix **zusammen**.*

- *beinhalten* (1 Vorkommen)

  (64) *Die Wandlungszeit der Analogausgabekanäle **beinhaltet** die Übernahme der digitalisierten Ausgabewerte und die Digital-Analog-Umsetzung.*

- *umfassen* (1 Vorkommen)

  (65) *Referenzdaten **umfassen** die Programmstruktur, die Querverweisliste, den Belegungsplan, die Liste nicht verwendeter Operanden und die Liste der Operanden ohne Symbol.*

Bei der Umfangsdefinition fand sich *bestehen aus* in acht und *enthalten* in drei Fällen. Je einmal erschienen Konstruktionen mit *sich zusammensetzen aus*, *beinhalten* und *umfassen*.

4.4.4 Programmbeschreibung

Nachfolgend sollen Ansatzpunkte zur Extraktion potenzieller Inhalts- und Umfangsdefinitionen aus dem erwähnten Handbuch beschrieben werden. Im Gegensatz zu den von Pearson und Meyer beschriebenen Vorgehen basiert die vorliegende Arbeit nicht auf einer vorgängigen Termerkennung, sondern wie bei Rebeyrolle qualifizieren alle Substantive als potenzielle Terme. Die Definitionskandidaten werden auf der Basis lexikalisch-syntaktischer Muster und unter Einbezug einiger ausschliessender Bedingungen extrahiert. Übernommen wurde der Ansatz von Pearson, das Verb als lexikalisches Muster zu verwenden und bezüglich der Art des den Begriff begleitenden Artikels Einschränkungen zu machen. Zudem wurde die Anregung von Meyer und Rebeyrolle aufgegriffen, durch Stopplisten ungültige Satzkandidaten auszuschliessen. Die Muster sind in der Programmiersprache Perl implementiert.

4.4.4.1 Vorbereitung

Da die Handbuchkapitel eine stark strukturierte Form mit zahlreichen Titeln aufweisen, wurde in einem ersten Schritt versucht, den zu betrachtenden Satz von Titeln oder anderen Informationen zu trennen. Da sich die Satzerkennung nach den entsprechenden typographischen Zeichen richtet, erschienen wie nachfolgend dargestellt Titel als Bestandteil des Satzes.

```
#BOS
7.10                CARD      --    NK      506
Bausteine           NN        --    --      0       %% Bau#stein
programmieren       VVFIN     --    --      0       %% programm~ier~en
Eine                ART       --    NK      503
SCL-Anweisung       NN        --    NK      503     %% SCL-An|weis~ung
ist                 VAFIN     --    --      0       %% sein
eine                ART       --    NK      501
ausführbare         ADJA      --    NK      501     %% aus|führ~bar
Aktion              NN        --    NK      501     %% Akt~ion
im                  APPRART   --    AC      500     %% in
Anweisungsteil      NN        --    NK      500     %% An|weis~ung\s#teil
eines               ART       --    NK      504
Codebausteins       NN        --    NK      504     %% Code#bau#stein
.                   $.        --    --      0
#EOS
```

Der Titel *Bausteine programmieren* verfälscht die Satzstruktur und erschwert die Anwendung von Mustern. Aus diesem Grund wurden sämtliche Zeilen vor dem letzten gross geschriebenen Artikel nicht berücksichtigt. Im obigen Beispiel resultiert dies im Satz *Eine SCL-Anweisung ist eine ausführbare Aktion im Anweisungsteil eines Codebausteins*. Einige Problemfälle konnten so gelöst werden. Bei der Extraktion zeigte sich allerdings, dass die Präsentation der Information in den Handbuchkapiteln ein Problem für die automatische Bearbeitung darstellen kann. So weist jedes Kapitel und Unterkapitel ein Inhaltsverzeichnis auf, zahlreiche Programmierbeispiele sind eingebaut und häufig werden Zusammenfassungen in stichwortartigen Aufstellungen gegeben. Da bei diesen Zusammenstellungen Satzendzeichen fehlen, wurden sie mit dem darauf folgenden Satz verbunden. Das führte in einigen Fällen dazu, dass gültige Sätze nicht erkannt wurden, wie in Abschnitt 4.4.5 näher diskutiert wird.

4.4.4.2 Inhaltsdefinition

Bei der Suche nach Inhaltsdefinitionen wird das Verb *sein* berücksichtigt. Als Bedingungen wurden folgende zwei Anforderungen formuliert:

- *Sein* muss als einziges Verb in einem Satzteil enthalten sein.
- Direkt vor und nach dem Verb *sein* muss eine Nominalphrase erscheinen.

Dieses Grundmuster deckt vier der für Inhaltsdefinitionen gefundenen Formen ab. Das Merkmal kann in diesem Muster durch einen Relativsatz, eine Präpositionalphrase, eine Nominalphrase im Genitiv oder eine Adjektivphrase ausgedrückt werden. Die so erhaltenen Ergebnisse zeigten, dass die von Pearson diskutierte Einschränkung der Art des begleitenden Artikels eine Verbesserung ergibt. Ausserdem wurden ungültige Sätze extrahiert, die Auflistungen mit Beispielcharakter enthalten. Sie waren häufig durch das Adjektiv *folgend* gekennzeichnet:

(66) *Die **folgenden** Beispiele sind zulässige Schreibweisen für eine Dezimalziffernfolge innerhalb von Literalen: […]*
(67) *Die **folgenden** Namen sind gültige Bezeichner: […]*

Aus diesen Beobachtungen ergaben sich zwei Einschränkungen:

- Nominalphrasen der Form Artikel + *folgend* + Substantiv werden ausgeschlossen.
- Ebenfalls werden Nominalphrasen mit dem Artikel *dies* nicht berücksichtigt, da dieser Artikel den Gültigkeitsbereich der Aussage einschränkt:

  (68) ***Dieser** Vereinbarungsblock ist Bestandteil des FB-Vereinbarungsteils.*

4.4.4.3 Umfangsdefinition

Die Basis der Extraktion von Kandidaten für Umfangsdefinitionen bilden die bei der manuellen Prüfung gefundenen Verben (*bestehen aus, beinhalten, umfassen, enthalten, sich zusammensetzen aus*). Die formulierten Bedingungen variieren je nach Verb leicht.

Die erste Gruppe umfasst die Verben *beinhalten, umfassen, enthalten*. Hier ist als einzige Bedingung gefordert, dass der Satz eines dieser Verben enthalten soll. Nach Durchsicht der Treffer wurde eine Ausschlussbedingung in Form einer Stoppliste formuliert. Metasprachliche Wörter wie *Tabelle, Abschnitt, Katalog, Handbuch, Kapitel, Bild* werden nicht zugelassen. Diese Wörter machen Aussagen zum Text selbst, z.B. *das folgende Kapitel enthält, die Tabelle 2.1 beinhaltet*.

Das zweite für die Umfangsdefinition untersuchte Muster ist das Verb *bestehen aus*. Auch hier gilt als einzige Bedingung, dass der Satz das Verb *bestehen* und eine durch *aus* eingeleitete Präpositionalphrase enthält. Die erzielten Resultate zeigten, dass einige ungültige Kandidaten Präpositionalphrasen mit dem Adjektiv *folgend* enthalten. Diese Struktur findet sich insbesondere bei der Beschreibung von Programmbeispielen:

(69) *Der Vereinbarungsteil besteht aus **folgenden** Teilen: […]*

Dabei handelt es sich allerdings um eine Eigenart des Testkorpus.

Als drittes Muster wurden Sätze mit dem Verb *zusammensetzen aus* gewählt. Dabei kommen zwei alternative Bedingungen zum Tragen:

- ein Satz enthält das Verb *zusammensetzen* oder
- ein Satzteil beinhaltet die Elemente *setzen* und *zusammen*.

Wie beim ersten Muster zeigte sich im Fall von *zusammensetzen*, dass der Ausschluss von metasprachlichen Wörtern die Genauigkeit des Rücklaufs erhöht.

4.4.5 Resultate

Die Resultate wurden auf der Basis von Recall und Precision formuliert. Recall (Vollständigkeit) und Precision (Genauigkeit) sind zwei Grössen, die das Mass der Effektivität einer automatisierten Suche bestimmen. Sie lassen sich wie folgt berechnen:[231]

$$\text{Recall} = \frac{\text{Anzahl der gefundenen relevanten Objekte}}{\text{Anzahl aller relevanten Objekte}}$$

$$\text{Precision} = \frac{\text{Anzahl der gefundenen relevanten Objekte}}{\text{Anzahl aller gefundenen Objekte}}$$

Der Recall-Wert sagt also etwas darüber aus, wie viele relevante Sätze gefunden wurden verglichen mit der Anzahl aller relevanten Sätze im Testkorpus. Die Precision setzt die Zahl der gefundenen gültigen Kandidaten ins Verhältnis zur Zahl der insgesamt gefundenen Sätze. Sie gibt an, wie viele der gefundenen Sätze relevant sind. Da das Testkorpus klein war, ist die Aussagekraft der nachfolgenden Prozentangaben zu Precision und Recall beschränkt.

[231] Salton 1989, S. 248

4.4.5.1 Precision

Die Genauigkeitswerte der verschiedenen Muster sind in Tabelle 4 aufgeführt.

Tabelle 4 Precision für die Muster zur Extraktion von Definitionskandidaten (Verhältnis zwischen relevanten und gefundenen Sätzen)

Muster	Precision	Anz. rel. Sätze/ Anz. gef. Sätze
NP *sein* NP	65%	37/57
beinhalten, umfassen, enthalten	33%	21/64
bestehen aus	65%	17/26
zusammensetzen aus	89%	8/9

Die Tabelle 4 zeigt schwankende Werte für die Genauigkeit. Sie liegen zwischen 33% für Sätze mit den Verben *beinhalten, umfassen, enthalten* und 89% für das Muster *zusammensetzen aus*. Die Suche nach Inhaltsdefinitionen und nach Umfangsdefinitionen der Form *bestehen* + *aus*-PP zeigen beide eine Precision von 65%. Aus diesen Werten lassen sich einige Schlüsse ableiten, vor allem zu den Mustern mit dem höchsten und tiefsten Prozentsatz.

Die geringste Precision weist das auf den Verben *beinhalten, umfassen, enthalten* basierende Muster auf. Gleichzeitig handelt es sich dabei um die produktivste Suche. 64 Kandidaten wurden gefunden, von denen aber nur 21 relevant sind. Bei der Durchsicht der Resultate zeigte sich, dass die grosse Menge an gefundenen Kandidaten auf das Verb *enthalten* zurückzuführen ist. Es erscheint in 50 Fällen, von denen nur 12 als gültige Definitionen gewertet wurden. *Beinhalten* und *umfassen* treten seltener auf: *Beinhalten* figuriert fünfmal, *umfassen* neunmal. Im Fall von *beinhalten* handelt es sich in vier Fällen um gültige Kandidaten, bei *umfassen* sind fünf der neun gefundenen Sätze definitorisch.

Die höchste Genauigkeit zeigt sich beim Muster *zusammensetzen aus*, bei dem es sich mit nur neun gefundenen Instanzen aber auch um das am wenigsten produktive handelt. Sätze mit dem Verb *zusammensetzen* erscheinen selten, konnten aber beinahe immer als Umfangsdefinitionen betrachtet werden. Es zeigte sich auch, dass eine explizite Suche nach einer durch *aus* eingeleiteten PP nicht notwendig ist. *Zusammensetzen* oder die Elemente *zusammen* und *setzen* traten ausschliesslich in Verbindung mit einer *aus*-PP auf.

In diesen beiden Fällen tritt die Verwendbarkeit der untersuchten Muster klar zutage. Beim ersten Muster sind restriktive Einschränkungen zu Sätzen mit dem Verb *enthalten* notwendig. So könnte beispielsweise zwingend eine durch *und* oder Komma signalisierte Aufzählung in Verbindung mit diesem Verb verlangt werden. Zu erwägen wäre auch, das Verb nicht zu verwenden. Im zweiten Fall zeigt sich, dass das Muster *zusammensetzen aus* nicht sehr produktiv ist, aber eine hohe Genauigkeit bietet.

Bei den Mustern mit mittleren Precision-Werten, NP *sein* NP und Sätze mit *bestehen aus*, gestalten sich Aussagen aufgrund der kleinen Grösse des Testkorpus schwierig. Im Fall von *bestehen aus* ist die Anzahl der gefundenen nicht relevanten Sätze mit neun Instanzen zu klein, als dass Rückschlüsse möglich wären. Bei der Inhaltsdefinition lassen sich zwei mögliche Verbesserungsansätze identifizieren:

- In drei Fällen erscheint *sein* in Verbindung mit einem Adjektiv, wie im folgenden Beispiel:

  (70) *Ein komplettes Feld ist einem anderen Feld **zuweisbar**, wenn sowohl die Datentypen der Komponenten als auch die Feldgrenzen (kleinst- und grösstmögliche Feldindizes) übereinstimmen.*

  Sein funktioniert hier als Kopula zu einem Adjektiv und verbindet nicht zwei Nominalphrasen. Wenn sich der Kasus der Nominalphrase bestimmen lässt, können diese Instanzen durch die Forderung nach einer Nominalphrase im Nominativ unterdrückt werden.
- In vier Sätzen finden sich die Ausdrücke *ist Bestandteil* bzw. *ist Inhalt*.[232] Anhand eines grösseren Korpus könnten Stopplisten ausgearbeitet werden, um Sätze wie den folgenden zu unterdrücken:

  (71) *Die Regeln sind **Bestandteil** der Sprachbeschreibung.*

4.4.5.2 Recall

Im Folgenden werden die Recall-Werte erörtert, d.h. das Verhältnis der gefundenen gültigen Kandidaten zu allen relevanten Sätzen im Korpus. Zur Berechnung des Recall wurden sämtliche Sätze ausgefiltert, die dem jeweils grundlegenden Erkennungskriterium entsprechen. Im Fall der Inhaltsdefinitionskandidaten ist dies die Bedingung, dass *sein* als einziges Verb in einem Satzteil erscheint. Bei der Umfangsdefinition wurden sämtliche

[232] Diese Ausdrücke können Meronymie-Beziehungen signalisieren. Sie traten in den im Testkorpus identifizierten Sätzen allerdings nicht in Verbindung mit Termen auf und waren daher nicht relevant.

Sätze extrahiert, die eines der als Muster verwendeten Verben enthalten. Die auf dieser Basis berechneten Recall-Werte sind in Tabelle 5 aufgeführt.

Tabelle 5 Recall für die Muster zur Extraktion von Definitionskandidaten (Verhältnis zwischen gefundenen relevanten und allen relevanten Sätzen)

Muster	Recall	Anz. gef. rel. Sätze/ Anz. rel. Sätze
NP *sein* NP	62%	37/60
beinhalten, umfassen, enthalten	87%	21/24
bestehen aus	100%	17/17
zusammensetzen	100%	8/8

Mit Ausnahme des Musters NP *sein* NP, bei dem sich die Recall- und Precisionwerte in etwa entsprechen, liegen die Recall-Prozentsätze höher als die Verhältniszahlen bei der Genauigkeit. Das Muster für die Inhaltsdefinition weist einen Wert von 62% auf und zeigt mit 60 Instanzen auch die grösste Zahl an definitorischen Sätzen. Umfangsdefinitionen sind mit insgesamt 49 Vorkommen seltener. Für die Verben *beinhalten*, *umfassen*, *enthalten* finden sich 24 definitorische Sätze, für *bestehen aus* 17 und für *zusammensetzen* 8. Bei den Umfangsdefinitionskandidaten liegen die Recall-Werte höher als bei der Inhaltsdefinition. Im Fall von zwei Mustern, *bestehen aus* und *zusammensetzen*, wurden alle definitorischen Sätze gefunden. Bei den Verben *beinhalten, umfassen, enthalten* liegt der Wert bei 87%.

In fünf Fällen führten Formatierungsprobleme zur Nichterkennung des Musters für Inhaltsdefinitionen. Aufstellungen wie Inhaltsverzeichnisse, Programmierbeispiele und stichwortartige Zusammenfassungen wurden irrtümlich dem nachfolgenden Satz zugeordnet, und damit war *sein* nicht das einzige Verb innerhalb eines Satzteils. Ein weiterer Grund für nicht gefundene Inhaltsdefinitionen ist der Einschub von Elementen nach dem Term. Dabei zeigten sich zwei typische Strukturen:

- NP_{Term}, *d.h. […], sein* $NP_{Hyperonym}$

  (72) *Die Zykluszeit**, d.h. die Zeit, bis ein Analogeingangswert wieder gewandelt wird,** ist die Summe der Wandlungszeiten aller aktivierten Analogeingabekanäle der Analogeinbaugruppe.*

  Ein durch *das heisst* eingeleiteter Einschub findet sich in sechs Fällen. Diese Struktur kann wie im obigen Beispiel auch im eingeschobenen Teil definitorische Informationen enthalten.

- NP_{Term} *(*Abkürzung/Vollbenennung*) sein* $NP_{Hyperonym}$

  (73) *Der Peripheriebus **(P-Bus)** ist ein serieller Rückwandbus […]*

  In drei Fällen wurde die Abkürzung oder Vollbenennung des Begriffs in Klammern nachgestellt.

Weiter zeigte sich ein bereits bei der manuellen Bestimmung von Definitionskandidaten identifiziertes Muster als Ursache für nicht extrahierte definitorische Sätze.

- NP_{Term} + *sein* + $NP_{Hyperonym}$ mit durch vorangestellte PP ergänztem partizipialem Adjektiv

  (74) *Die STEP7-Bausteine sind **durch ihre Funktion, ihre Struktur oder ihren Verwendungszweck abgegrenzte** Teile eines Anwenderprogramms.*

  Wenn das Merkmal durch ein partizipiales Adjektiv realisiert ist, das durch eine vorangestellte PP ergänzt wird, wurde es mit dem verwendeten Suchmuster nicht extrahiert.

4.4.6 Manuelle Prüfung von Begriffsumgebungen zur Identifikation weiterer definitorischer Formen

Als letzter Projektschritt wurde eine manuelle Prüfung von Termkontexten durchgeführt. Das Ziel war zum einen, die Häufigkeit verschiedener Definitionsarten zu evaluieren. Weiter sollten Ansätze für die Extraktion von Begriffsbestimmungen geprüft werden, die nicht der Form einer Inhalts- oder Umfangsdefinition entsprechen. Diese Bestimmung basiert auf 500 Sätzen, die eine in der SIMATIC-TermDB aufgeführte Benennung enthalten. 60 Sätze wurden als definitorisch relevant eingestuft. Folgende Typen lassen sich unterscheiden:

- Definition durch den Verbinhalt (18 Vorkommen)

  (75) *Der RS 485-Repeater **verstärkt** Datensignale auf Busleitungen und koppelt Bussegmente.*

  Am häufigsten finden sich Definitionen durch Funktionsbeschreibungen. In diesen Fällen stellt üblicherweise das Verb das bedeutungstragende Element dar. Sie können daher nicht systematisch erkannt werden und sind für eine automatische Identifikation kaum geeignet.

- Sätze in der Form einer Inhalts- oder Umfangsdefinition (11 Vorkommen)

  Hier zeigte sich als neues potenzielles Verb für eine Umfangsdefinition *unterteilt sein*.

  (76) *Der P-Bus **ist** in zwei P-Bussegmente mit 10 bzw. 8 Steckplätzen **unterteilt**.*

- Satzzeichen und Formatierung (11 Vorkommen)

  In 6 Fällen ist die Benennung in Klammern enthalten, während der Satz eine Erklärung liefert.

  (77) *Achten Sie auf die Aussparung an der Vorderseite des Moduls **(Verpolschutz)**.*

  In 5 Fällen sind Aufzählungen auszumachen, die entweder die Teile eines Begriffs auflisten oder Ober- und Unterbegriff kennzeichnen.

  (78) *Für Signalbaugruppen der S7-400 gibt es 3 Typen von Frontsteckern**:***
  - Frontstecker mit Crimpanschluss
  - Frontstecker mit Schraubanschluss
  - Frontstecker mit Federkraftanschluss

- Präpositionen und Konjunktionen (10 Vorkommen)

  Einige Präpositionen und Konjunktionen können definitorische Erklärungen einleiten. Zu den zwei häufigsten gehörten *zu* und *um […] zu*. Sie erläutern üblicherweise die Funktion eines Begriffs:

  (79) *Filtermatte (optional): **Zur** Filterung der Zuluft können Sie bei Kabelkanal und Lüfterzeile eine Filtermatte einbauen.*
  (80) ***Um** eine Baugruppe mit ihrer Steckplatznummer **zu** kennzeichnen, verwenden Sie Steckplatzschilder.*

  In je einem Fall fanden sich *wenn […] dann* und *damit*.

- Verben und Wendungen (10 Vorkommen)

  Das Verb *dienen zu* signalisiert häufig eine Funktionsbeschreibung.

  (81) *Das äußere Schirmgeflecht **dient zur** Ableitung von Störströmen und muss in die Schirmungs- und Erdungsmaßnahmen einbezogen werden.*

  Erläuterungen finden sich auch bei den folgenden Verben und Wendungen: *Eigenschaften haben, die Möglichkeit bieten, bestimmt werden durch, entstehen bei* und *das heisst*.

Bei einer weitergehenden Untersuchung wäre es lohnend, die Häufigkeit von durch Satzzeichen und Formatierungen gekennzeichneter Information zu ermitteln. Ebenso könnte die Liste der lexikalischen Muster erweitert und anhand eines grösseren Korpus auf deren Nützlichkeit geprüft werden.

4.4.7 Zusammenfassung und weiterführende Arbeiten

Bei der Wertung der Resultate muss die beschränkte Grösse des Testkorpus beachtet werden. Die zur Extraktion von Definitionskandidaten verwendeten Muster waren offen formuliert und benutzten nur wenige Einschränkungen. Die erzielten Recall-Werte für Umfangsdefinitionen waren entsprechend hoch. Die Vollständigkeit der einer Inhaltsdefinition entsprechenden Begriffsbestimmungen könnte durch die Implementierung eines zusätzlichen Musters verbessert werden, bei dem das Merkmal durch ein partizipiales Adjektiv, ergänzt durch eine Präpositionalphrase, ausgedrückt wird. Weiter ist die Möglichkeit von Einschüben, die dem Begriff folgen, zu berücksichtigen. Die Untersuchung der Precision zeigte, dass sich das Verb *zusammensetzen* besonders gut für Umfangsdefinitionen eignet, während das Verb *enthalten* zahlreiche ungültige Kandidaten produziert. *Bestehen aus*, *beinhalten* und *umfassen* weisen mittlere Precision-Werte auf. Diese Verben müssen an einem grösseren Korpus getestet werden, um verlässliche Aussagen machen zu können. Im Fall des Musters NP *sein* NP fanden sich zwei mögliche Verbesserungsansätze. Stopplisten und der Ausschluss von *sein* als Kopula zwischen Nominalphrase und Adjektiv könnten die Genauigkeit verbessern.

Des Weiteren wäre eine Untersuchung von lexikalischen Elementen lohnend, die in Verbindung mit Begriffen auf definitorische Informationen hinweisen. Die manuelle Prüfung zeigte, dass im Testkorpus vor allem das Verb *dienen zu* häufig Information zu Funktion und Zweck eines Begriffs liefert. Auch die Zuhilfenahme von Satzzeichen und Formatierungen bietet einen Ansatzpunkt zur Identifikation von Begriffserklärungen.

5 ZUSAMMENFASSUNG

Ziel der vorliegenden Arbeit war die Betrachtung der terminologischen Definition unter theoretischen und praktischen Gesichtspunkten sowie ihre automatische Extraktion. Kapitel 2.1 beschäftigte sich mit der Sicht der Terminologielehre und präsentierte die grundlegenden Aussagen zu Definitionen. Als Erstes wurden die zwei für die Terminologie relevantesten Definitionsformen, die Inhalts- und die Umfangsdefinition, erörtert. Sie stellen die Ordnung des Begriffssystems explizit dar und sind daher für die terminologische Tätigkeit geeignet. Aus praktischen Gründen nennt die Literatur weitere Definitionsformen. Deren Zusammenstellung fand sich in Abschnitt 2.1.2, gefolgt von einer Darstellung der an Definitionen gestellten Anforderungen inhaltlicher und formaler Art. Hier zeigte sich, dass die Anforderungen unterteilt werden können in Bedingungen, die bei sämtlichen terminologischen Tätigkeiten realisierbar sind, und solche, die nur im Rahmen der normenden Arbeit umgesetzt werden können. Eine Unterscheidung zwischen normender, deskriptiver und übersetzungsorientierter Tätigkeit wird von zahlreichen Autoren auch bei der Bestimmung des Stellenwerts der Definition gemacht. Die Relevanz von terminologischen Definitionen bildete das Thema von Kapitel 2.2. Die Literatur ist sich einig, dass die Definition bei der normenden Tätigkeit zwingend, bei der übersetzungsorientierten Arbeit fakultativ ist. Weiter findet sich die Meinung, dass Definitionen ohne Informationsverlust durch Kontexte ersetzt werden können. Im Rahmen einer Befragung der Übersetzer eines Sprachendiensts zeigte sich allerdings, dass die Definition auch bei der übersetzungsorientierten Arbeit als relevante Datenkategorie eingestuft wird und dass ein Bedürfnis nach Begriffsbestimmungen besteht. Diese Auffassung wurde durch die in einem Terminologieprojekt gemachten Erfahrungen bestätigt. Der Informationsgehalt von Einträgen ohne Definitionen ist stark reduziert, und Bereinigungsarbeiten werden erschwert oder verunmöglicht. Im Rahmen des Projekts zeigte sich auch, dass einige der von der Literatur formulierten Anforderungen in der übersetzungsorientierten Arbeit nicht umsetzbar sind, da ein abgegrenztes und zusammenhängendes Begriffssystem häufig nicht gegeben ist.

Da die Definition auch für die übersetzungsgerichtete Arbeit wichtig ist, aus Zeitgründen jedoch häufig vernachlässigt wird, bietet sich eine automatische Extraktion von Definitionskandidaten an. Sie erleichtert auch die Anpassung von Definitionen an sich ver-

ändernde Begriffsinhalte. Kapitel 4 beleuchtete verschiedene Ansätze zur Erkennung kontextualisierter Definitionen. Unterscheiden lassen sich Verfahren, die die Extraktion definitorischer Sätze an sich bezwecken und solche, die den Schwerpunkt auf die strukturierte Darstellung der Bestandteile einer Definition legen. Weiter existieren diesem Thema verwandte Arbeiten, die sich mit der Identifikation von Begriffsbeziehungen befassen. Ein Grossteil der Studien untersucht die Extraktion von englischen Definitionskandidaten. In Kapitel 4.4 wurden Bedingungen für die Erkennung deutscher Definitionskandidaten vorgestellt. In diesem Kapitel findet sich auch eine Darstellung der manuell identifizierten Erscheinungsformen von Inhalts- und Umfangsdefinitionen sowie weiterer Definitionsarten. Dabei zeigten sich Möglichkeiten zur Verbesserung der implementierten Kriterien und Ansatzpunkte zur Identifikation anderer Formen definitorischer Sätze.

Verfahren zur automatischen Identifikation definitorischer Sätze sind in verschiedenen Gebieten einsetzbar. Eine offensichtliche Anwendung liegt im Bereich der Terminographie. Hier können Einträge mit wichtigen terminologischen Informationen ergänzt werden. Weiter ist eine Verbindung mit Systemen zur Termextraktion vorstellbar. Erscheinen Wörter in definitorischen Sätzen, erhöht sich die Wahrscheinlichkeit, dass es sich dabei um Terme handelt. Verfahren zur Definitionsidentifikation können daher die Termerkennung unterstützen. Ein weiteres mögliches Anwendungsgebiet liegt in der Analyse von Texten zum Aufbau von Korpora. Aussagekräftige Texte weisen eine gewisse Anzahl an definitorischen Sätzen auf, und ein automatisches Verfahren zu deren Erkennung kann als Kriterium für die Auswahl von Texten dienen. Zusammenfassend kann die Extraktion von Definitionskandidaten als lohnendes Forschungsgebiet im Bereich der automatischen Sprachverarbeitung bezeichnet werden.

6 BIBLIOGRAPHIE

Ahmad, K. und H. Fulford. 1992. *Knowledge Processing 4. Semantic Relations and their Use in Elaborating Terminology.* Computing Sciences Report CS-92-07. Guildford: University of Surrey.

Arntz, Reiner und Heribert Picht. 1995. *Einführung in die Terminologiearbeit*. Studien zu Sprache und Technik 2. Hildesheim: Georg Olms.

Aussenac-Gilles, Nathalie und Patrick Séguéla. 2000. "Les relations sémantiques: du linguistique au formel". *Cahiers de Grammaire* 25, 175-201.

Austin, J.J. 1962. *How to do things with words.* Oxford: Oxford University Press.

Borillo, Andrée. 1996. "Exploration automatisée de textes de spécialité: repérage et identification automatique de la relation lexicale d'hyperonymie". *Linx* 34-35, 113-121.

Bowden, P.R., P. Halstead und T.G. Rose. 1996. "Extracting Conceptual Knowledge From Text Using Explicit Relation Markers". In: N. Shadbolt, K. O'Hara und G. Schreiber (Hrsg.). *Advances in Knowledge Acquisition. Proceedings of the 9th European Knowledge Acquisition Workshop, EKAW'96.* Nottingham, 147-162.

Bramki, D. und R. Williams. 1984. "Lexical familiarization in economics text books". *Reading in a Foreign Language* 2/1, 169-181.

Büchel, Gregor und Nico Weber. 1995. "Semantische Relationen in Definitionsstrukturen". In: L. Hitzenberger (Hrsg.). *Angewandte Computerlinguistik: Vorträge im Rahmen der Jahrestagung 1995 der Gesellschaft für Linguistische Datenverarbeitung (GLDV) e.V.* Hildesheim: Olms, 127-140.

Büchel, Gregor. 1996. "Können Verben semantische Relationen markieren?". In: N. Weber (Hrsg.). *Semantik, Lexikographie und Computeranwendungen*. Tübingen: Niemeyer, 275-286.

Chaudron, C. 1982. "Vocabulary elaboration in teachers' speech to L2 learners". *Studies in Second Language Acquisition* 4/2, 170-180.

Condamines, Anne und Josette Rebeyrolle. 2001. "Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB)". In: D. Bourigault, Ch. Jacquemin und M.-C. L'Homme (Hrsg.). *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins, 127-148.

Dahlberg, Ingetraut. 1981. "Conceptual definitions for INTERCONCEPT". *International Classification* 8/1, 16-22.

De Bessé, Bruno. 1990. *Stage de terminologie. Notes de cours*. Genf.

De Bessé, Bruno. 1997. "Terminological Definitions". In: S.E. Wright und G. Budin (Hrsg.). *Handbook of Terminology Management.* Bd. I. Amsterdam: John Benjamins, 63-74.

DIN 2330. 1979. *Begriffe und Benennungen: Allgemeine Grundsätze*. Berlin, Köln: Beuth.

DIN 2342 Teil 1. Entwurf 1986. *Begriffe der Terminologielehre. Grundbegriffe*. Berlin, Köln: Beuth.

DIN 2342 Teil 1. 1992. *Begriffe der Terminologielehre. Grundbegriffe*. Berlin, Köln: Beuth.

Drozd, Lubomir und W. Seibicke. 1973. *Deutsche Fach- und Wissenschaftssprache. Bestandsaufnahme – Theorie – Geschichte*. Wiesbaden: Brandstetter.

Drozd, Lubomir. 1983. "Terminological Synonyms and the Function of Definition: Theses". In: D. Ducet-Picard (Hrsg.). *Problèmes de la définition et de la synonymie en terminologie. Actes du Colloque International de Terminologie, GIRSTERM*. Québec: Université Laval, 87-100.

Felber, Helmut. 1990. *Die Lehre von den allgemeinen terminologischen Grundsätzen und Methoden.* Wien: TermNet.

Felber, Helmut. 1993. *Allgemeine Terminologielehre und Wissenstechnik – theoretische Grundlagen*. Wien: TermNet.

Flowerdew, John. 1992a. "Definitions in Science Lectures". *Applied Linguistics* (13)2, 202-221.

Flowerdew, John. 1992b. "Salience in the Performance of One Speech Act: The Case of Definitions". *Discourse processes* 15, 165-181.

Harris, Brian M. 1983. "The need for definitions on term records: a translator's view". In: D. Ducet-Picard (Hrsg.). *Problèmes de la définition et de la synonymie en terminologie. Actes du Colloque International de Terminologie, GIRSTERM*. Québec: Université Laval, 141-156.

Hearst, Marti A. 1992. "Automatic Acquisition of Hyponyms from Large Text Corpora". In: Ch. Boitet (Hrsg.). *Proceedings of the international conference on computational linguistics (COLING-92)*. Nantes, 539-545.

Hohnhold, Ingo. 1982. "Grundbegriffe im Bereich und im Umfeld übersetzungsorientierter Terminologiearbeit. Arbeitsdefinitionen und Anmerkungen". *Lebende Sprachen* 27, 1-5.

IBM. 1985. *Fachausdrücke der Informationsverarbeitung.* Stuttgart: Klett.

ISO 704. 1987. *Principles and methods of terminology.* Genf: International Organization for Standardization.

ISO/DIS 1087. Entwurf 1988. *Terminology – Vocabulary. Revision of ISO/R 1087: 1969*. Genf: International Organization for Standardization.

ISO/R 1087. 1969. *The Vocabulary of Terminology*. Genf: International Organization for Standardization.

Jouis, Christophe. 1993. *Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes. Réalisation d'un prototype: le système SEEK.* Diss., EHESS, Paris.

KÜWES (Konferenz der Übersetzungsdienste westeuropäischer Sprachen). 1990. *Empfehlungen für die Terminologiearbeit*. Bern.

Lambrou, A.V. 1979. *Definitions in Undergraduate Science Text Books*. Diss., University of Khartoum, Sudan.

Landau, Sidney. 2001. *Dictionaries: The Art and Craft of Lexicography.* 2. Auflage. Cambridge: Cambridge University Press.

Le Priol, Florence. 2000. *Extraction et capitalisation automatiques de connaissances à partir de documents textuels. SEEK-JAVA: identification et interprétation de relations entre concepts.* Diss., Université Paris-Sorbonne.

Lingsoft-Oy. 1994. "Gertwol. Questionnaire for Morpholympics 1994". *LDV-Forum* 11(1), 17-29.

Mayer, Felix. 1998. *Eintragsmodelle für terminologische Datenbanken. Ein Beitrag zur übersetzungsorientierten Terminographie.* Tübingen: Gunter Narr.

Meyer, Ingrid. 2001. "Extracting knowledge-rich contexts for terminography. A conceptual and methodological framework". In: D. Bourigault, Ch. Jacquemin und M.-C. L'Homme (Hrsg.). *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins, 279-302.

Microsoft Press. 2001. *COMPUTER. Lexikon, Fachwörterbuch.* Unterschleissheim: Microsoft.

Miller, George A. 1999. "Nouns in WordNet". In: Ch. Fellbaum (Hrsg.). *WordNet. An Electronic Lexical Database.* Cambridge: MIT, 23-46.

Morin, Emmanuel und Emmanuelle Martienne. 1999. "Raffinement de patrons lexico-syntaxiques par un système d'apprentissage". In: R. Teulier-Bourgine (Hrsg.). *Actes, Troisième Conférence sur l'Ingénierie des Connaissances (IC'99).* Palaiseau, 89-96.

Natanson, Edouard. 1983. "Rapports entre la définition et la délimitation des concepts". In: D. Ducet-Picard (Hrsg.). *Problèmes de la définition et de la synonymie en terminologie. Actes du Colloque International de Terminologie , GIRSTERM*. Québec: Université Laval, 55-65.

Ndi-Kimbi, Augustin. 1994. "Guidelines for terminological definitions: The adherence to and deviation from existing rules in BS/ISO 2382: Data Processing and Information Technology Vocabulary". *Terminology* 1(2), 327-350.

ÖNORM 2710. Entwurf 1993. *Übersetzungsorientierte Terminographie. Terminographische Datenkategorien und Richtlinien für ihre Anwendung.* Wien.

Otman, Gabriel. 1996. "Le traitement automatique de la relation partie-tout en terminologie." *Faits de langue* 7, 43-52.

Pearson, Jennifer. 1998. *Terms in Context.* Amsterdam: John Benjamins.

Rebeyrolle, Josette. 2000. "Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes". In: P. Tchounikine (Hrsg.). *Actes Journées Francophones d'Ingénierie des Connaissances, IC'2000.* Toulouse, 105-114.

Riloff, Ellen. 1996. "Automatically Generating Extraction Patterns from Untagged Text". In: *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96).* Portland, 1044-1049.

Rousseau, Louis-Jean. 1983. "La définition terminologique". In: D. Ducet-Picard (Hrsg.). *Problèmes de la définition et de la synonymie en terminologie. Actes du Colloque International de Terminologie, GIRSTERM*. Québec: Université Laval, 35-46.

Rousselot, F., P. Frath und R. Oueslati. 1996. "Extracting concepts and relations from corpora". In: W. Wahlster (Hrsg.). *Proceedings of ECAI'96, 12th European Conference on Artificial Intelligence*. Budapest, 74-78.

Sager, Juan C. 1983. "Definitions in Terminology". In: D. Ducet-Picard (Hrsg.). *Problèmes de la définition et de la synonymie en terminologie. Actes du Colloque International de Terminologie, GIRSTERM*. Québec: Université Laval, 113-139.

Sager, Juan C. 1990. *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins.

Sager, Juan C. und Augustin Ndi-Kimbi. 1995. "The conceptual structure of terminological definitions and their linguistic realisations: A report on research in progress". *Terminology* 2(1), 61-81.

Salton, G. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Reading: Addison-Wesley.

Schmid, H. und A. Kempe. 1996. "Tagging von Korpora mit HMM, Entscheidungsbäumen und Neuronalen Netzen". In: H. Feldweg und E.W. Hinrichs (Hrsg.). *Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschliessung des Deutschen.* Niedermeyer: Tübingen, 231-244.

Schneider, Hans-Jochen. 1997. *Lexikon Informatik und Datenverarbeitung*. Frankfurt: Oldenbourg.

Skut, W., B. Krenn, T. Brants und H. Uszkoreit. 1997. "An annotation scheme for free word order languages". In: Association for Computational Linguistics (Hrsg.). *Proceedings of the 5th Conference on Applied Natural Language Processing*. Washington D.C., 88-95.

Strehlow, R.A. 1983. "Terminology and the well-formed definition". In: C.G. Interrante und F.J. Heymann (Hrsg.). *Standardization of Technical Terminology: Principles and Practices*. Philadelphia: American Society for Testing and Materials, 15-25.

Thielen, C. und A. Schiller. 1996. "Ein kleines und erweitertes Tagset fürs Deutsche". In: H. Feldweg und E.W. Hinrichs (Hrsg.). *Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschliessung des Deutschen.* Niedermeyer: Tübingen, 193-203.

Trimble, L. 1985. *English for Science and Technology: A Discourse Approach*. Cambridge: Cambridge University Press.

Volk, Martin. 2001. *The Automatic Resolution of Prepositional Phrase-Attachment Ambiguities in German*. Hab., Philosophische Fakultät Universität Zürich.

Weber, Nico. 1993. "Computergestützte Analyse von Definitionstexten in einem deutschen Wörterbuch". In: Horst P. Pütz und Johann Haller (Hrsg.). *Sprachtechnologie: Methoden, Werkzeuge, Perspektiven: Vorträge im Rahmen der Jahrestagung 1993 der Gesellschaft für Linguistische Datenverarbeitung (GLDV) e.V.* Hildesheim: Olms, 140-168.

Winkler, Peter. 2000. *Computerlexikon, M+T*. München: Wilhelm Heyne.

Wright, Sue Ellen. 2001. "Punktuelle Terminologiearbeit in modernen Übersetzungsumgebungen". *Mitteilungen für Dolmetscher und Übersetzer* 47/1, 5-10.

Wüster, Eugen. 1991. *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. 3. Auflage. Bonn.