aus: Carstensen, K.-U., Ebert, C., Ebert, C., Jekat, S., Klabunde, R. und Langer, H., Hrsg. (2009): Computerlinguistik und Sprachtechnologie – Eine Einführung. 3. überarbeitete und erweiterte Auflage, Spektrum Akademischer Verlag.

1 Nicht-sprachliches Wissen

Kai-Uwe Carstensen

1.1 Die Relevanz nicht-sprachlichen Wissens für die CL

Nicht-sprachliches Wissen ist sicherlich kein zentraler Untersuchungsgegenstand der Computerlinguistik im Allgemeinen. Auch für die Linguistik scheint es nur insofern relevant zu sein, als sprachliche Zeichen grundsätzlich sowohl eine Form- wie auch eine –nicht-sprachliche– Inhaltsseite aufweisen und die Modellierung der einen Seite somit von den Erkenntnissen über die andere profitiert (dies gilt für beide Richtungen, s. hierzu als Beispiel Lang, Carstensen, and Simmons, 1991). In der formalen Semantik war nicht-sprachliches Wissen lange Zeit als sog. „Weltwissen“ verpönt, das keinen sprachwissenschaftlichen Mehrwert darstellt.

Tatsächlich ist die aus der Vernachlässigung der Wortsemantik resultierende Gleichsetzung nicht-sprachlichen Wissens mit Weltwissen aber unangemessen: nicht-sprachliches Wissen umfasst nicht nur die Kenntnis, was der Fall ist (Faktenwissen, episodisches Wissen), sondern ebenfalls, was es überhaupt gibt/wie unsere Auffassung der Welt strukturiert ist (konzeptuelles Wissen). Angelehnt an die philosophische Teildisziplin **Ontologie** („Lehre vom Seienden “) spricht man daher bei Letzterem auch von **ontologischem Wissen**. Moderne semantische Ansätze berücksichtigen diesen Umstand (vgl. die Qualiastruktur in Unterkapitel **??** oder die „ontological semantics“ von Nirenburg and Raskin 2004).

Nicht-sprachliches Wissen ist außerdem zentraler Bestandteil intelligenter Systeme (daher: „wissensbasierte“ Systeme). Es wird zur Kategorisierung sensorischen Inputs („Daten“), zur Problemlösung, zur Handlungsplanung und Kommunikation benötigt. *Konzeptuelle* Repräsentationen vermitteln zwischen Wahrnehmung, Handlung und Sprache.

In frühen Anwendungssystemen der Künstlichen Intelligenz (KI), den Expertensystemen, führte das Fehlen von Wissen über die Welt zur so genannten „Zerbrechlichkeit“ (*brittleness*): Schon die geringsten Abweichungen von vorgegebenen Eingabemustern führ(t)en zu Fehlern und Systemabstürzen, die für die Benutzer nicht nachvollziehbar waren. Hieraus enstand das Desiderat allgemein und wieder- verwendbarer Wissensressourcen (sog. *(common sense) knowledge sharing and reuse*), heute allgemein **Ontologien** genannt.

Für **Anwendungen** der Computerlinguistik ist nicht-sprachliches Wissen daher essentiell: im Verlauf des *Textverstehens* müssen Textrepräsentationen mit Hintergrundwissen verrechnet werden (z. B. für die Auflösung von Ambiguitäten und bei der Präsuppositionsrechtfertigung); nicht-sprachliche Wissensstrukturen sind Grundlage und Ausgangspunkt für die *Sprachgenerierung* (s. Unterkapitel **??**); in der *maschinellen Übersetzung* (s. Unterkapitel **??**) werden konzeptuelle Repräsentationen als Interlingua verwendet. Insbesondere basieren zukunftsweisende Versionen des World Wide Web wie z. B. das **Semantic Web** (s. Berners-Lee et al. 2001) auf Ontologien.

1.2 Was ist „Wissen“ (nicht)?

Was eigentlich ist **„Wissen“**: die Menge der Daten, auf die ein (natürliches oder künstliches) System zugreifen kann, die Menge an Information, die ihm zur Verfügung steht, die Menge der von ihm begründeten Annahmen, die wahr sind? Keine von diesen knappen Charakterisierungen ist zutreffend.

Zunächst einmal ist Wissen abstrakt und nicht in Form konkreter **Daten** zu erfassen; es ist etwas, das einem System zugeschrieben werden kann, ohne auf die konkrete Form zu verweisen, in der es realisiert ist. Zudem ist Wissen unendlich und zeigt so das Vorhandensein sowohl von Struktur- als auch von Verarbeitungsaspekten: Wissen liegt nicht nur explizit vor, sondern kann durch Inferenzprozesse erschlossen werden („Ich weiß, dass du weißt, dass ich weiß. . . dass ich existiere“). Gleichwohl bilden Daten eine wichtige Grundlage für Wissen, da Wissensstrukturen zum Teil anhand des systematischen Auftretens von Daten konstruiert werden (→ Lernen).

Wissen ist ebenfalls nicht gleichzusetzen mit **„Information“**, auch wenn das Paradigma der Informationsverarbeitung zentral für diesen Bereich ist. Information ist vielmehr das Bindeglied zwischen Daten und den Strukturen, die abstraktes Wissen realisieren: Daten sind dann Information, wenn sie als Instanzen schematischer Strukturen erkannt werden (dies charakterisiert die semantische Auffassung von „Information“ innerhalb der Kognitionswissenschaft, die von der syntaktischen, inhaltsleeren Auffassung der Informationstheorie zu unterscheiden ist). Entsprechend wird deutlich, dass beispielsweise **Informationsextraktion** (s. Unterkapitel **??**) als wichtige Anwendung der Computerlinguistik mehr ist als reines Sammeln von Daten, indem es zwingend Wissen voraussetzt. Gleichzeitig würde „Wissensextraktion“ implizieren, dass zusätzlich Wissenstrukturen aufgebaut werden.

Im Bereich der Computerlinguistik und der Künstlichen Intelligenz wird im Gegensatz zur philosophischen Tradition ein erweiterter Wissensbegriff verwendet (∼ **Kenntnis**): so ist z.B. „syntaktisches Wissen“ nicht mithilfe von begründeten wahren Annahmen oder ähnlichen, auf rationalen Erwägungen basierenden Konzepten zu beschreiben. Auch hier zeigt sich *Wissen* als abstrakte Beobachterkategorie, wodurch offen bleibt, wie es realisiert ist.

1.3 Wissen und Wissensrepräsentation

Kerngebiet der Beschäftigung mit **nicht-sprachlichem Wissen** ist der Bereich der **Wissensrepräsentation** innerhalb der KI bzw. der Kognitionswissenschaft. Interessanterweise waren es vor allem sprachlich orientierte Ansätze, die die Notwendigkeit und das Potential der Repräsentation nicht-sprachlichen Wissens aufgezeigt haben (z. B. Quillian 1968; Schank 1975), sowie das bekanntberüchtigte ELIZA-Programm Joseph Weizenbaums als Karikatur eines Systems, das gerade nicht über Wissensrepräsentationen verfügt.

„Wissensrepräsentation“ bezeichnet einerseits die Realisierung abstrakten Wissens in einem konkreten (physikalischen) System (also Mensch oder Maschine) und andererseits die formal zu erfassenden Strukturen, die sich aus der

Interaktion eines vorstrukturierten informationsverarbeitenden Systems mit seiner Umwelt (→ Lernen) ergeben. Das Verhältnis von Wissen zu dessen Repräsentation ist am markantesten in der sogenannten *Knowledge Representation Hypothesis* ausgedrückt:

> Any mechanically embodied intelligent process will be comprised of structural ingredients that a) we as external observers naturally take to represent a propositional account of the knowledge that the overall process exhibits, and b) independent of such external semantical attribution, play a formal but causal and essential role in engendering the behavior that manifests that knowledge. (Smith 1982, S.33).

Den „structural ingredients" entsprechen in *symbolischen Ansätzen* zur Wissensrepräsentation Systeme von Symbolen mit einem jeweils spezifischen Bedeutungsgehalt, wobei die Symbole als in irgendeiner Weise physikalisch realisiert aufgefasst werden (sog. *Physical symbol system hypothesis*, s. Newell and Simon 1976). Ein Beispiel hierfür sind Symbole für Konzepte, denen in der realen Welt Einzeldinge oder Mengen solcher Dinge entsprechen. Die Frage jedoch, wie diese Symbolsysteme genau in den Erfahrungen bzgl. der Umwelt verankert sind, ist als das „symbol grounding problem" (Harnad 1990) bekannt geworden.

Dem gegenüber stellen *subsymbolische* bzw. *konnektionistische Ansätze* das Verkörpertsein (embodiment) und die Situiertheit neuronaler Netze, die von vornherein auf einer Sensor/Input-Aktor/Output-Korrelation beruhen und in denen sich ein von außen zugeschriebener Bedeutungsgehalt (z. B. „Dies ist das Konzept für X") aus den Aktivationsmustern einer Vielzahl von Neuronen/Einheiten ergibt.

Beide Ansätze haben ihre Stärken und Schwächen und sind gegenwärtig daher als komplementär zueinander anzusehen (s. aber das in Hitzler and Kühnberger 2009 formulierte Desiderat der Synthese beider Bereiche). Beispielsweise erfassen subsymbolische Ansätze sehr viel besser die graduellen Unterschiede, kontextuellen Abhängigkeiten und impliziten Zusammenhänge in eng begrenzten Anwendungsbereichen. Symbolische Ansätze bleiben jedoch vorerst für die Entwicklung komplexer natürlich-sprachlicher Systeme besser geeignet. Dies gilt insbesondere für die Erstellung umfangreicher Ressourcen nicht-sprachlichen Wissens.

1.4 Aspekte der Wissensrepräsentation

1.4.1 Allgemeine Aspekte

Kern der Wissensrepräsentation ist die Darstellung der im Folgenden aufgeführten generischen Wissensrepräsentationskonstrukte: **Konzepte** (dt. Begriffe) als Repräsentanten von Entitäten der Welt (zu unterscheiden sind hier Klassenkonzepte (→ generisches Wissen über Dinge) und Individuenkonzepte/**Instanzen** (→ Wissen über Einzeldinge); **Attribute** als Repräsentanten der Eigenschaften solcher Entitäten; **Relationen** als Repräsentanten von Beziehungen zwischen Dingen; **Regeln** als Repräsentanten der Beziehungen zwischen Sachver-

halten. Die Aufgabe der Wissensrepräsentation ist die formale Explikation dieser Aspekte, so dass alles relevante Wissen —je nach Anspruch eingeschränkt auf bestimmte Bereiche (**Domänen**) oder Verwendungszwecke— entweder direkt repräsentiert ist oder anhand von Schlussfolgerungen (**Inferenzen**) systematisch erschlossen werden kann. Hierbei stellen sich viele Detailfragen (z. B.: Welche Repräsentationskonstrukte entsprechen dem Ausdruck „ist ein“?), deren Beantwortung die Kenntnisse der kognitionswissenschaftlichen Disziplinen (u.a. Informatik, Linguistik, Psychologie, Philosophie) erforderlich macht.

1.4.2 Paradigmen

Verschiedene Sichtweisen darauf, wie sich aus solchen allgemeinen Wissensrepräsentationskonstrukten konkrete Wissensrepräsentationsstrukturen konstruieren lassen (und welche Prozesse darüber ablaufen sollen) haben zu unterschiedlichen Paradigmen der Wissensrepräsentation geführt.

Den **semantischen Netzwerken** liegt die Idee der kognitiven Vernetztheit konzeptuellen Wissens zugrunde. Ihren Ursprung hat diese Auffassung in den Arbeiten Quillians (z. B. Quillian 1968), der entsprechende Repräsentationen zur Berechnung der inhaltlichen Beziehung sprachlicher Elemente (daher: „semantische“ Netzwerke) verwendete. Den „Knoten“ („nodes“) des semantischen Netzwerks entsprechen die Konzepte, den „Kanten“ („links“) die vielfältigen Beziehungen zwischen ihnen. Entsprechend lassen sich in einem solchen Netzwerk „Nähe“ bzw. „Ferne“ von Konzepten über die Länge der Pfade verbindender Kanten verstehen und auch psychologisch relevante Prozesse wie „Aktivationsausbreitung “ definieren.

Das **Frame-Paradigma** betont den objekt-orientierten und schematischen Aspekt der Wissensrepräsentation, wonach das relevante Wissen über eine Entität direkt an ihrem Stellvertreter verfügbar ist. Zentrales Konstrukt dieses Paradigmas ist das des Frames („Rahmen“, s. Minsky 1975) als Repräsentation schematischen Wissens über Entitäten der Welt (Stühle, Kindergeburtstage etc.). Frames sind im Wesentlichen Attribut-Wert-Paare, wobei die Werte („Filler“) der Attribute („Slots“) bei Fehlen genauerer Information durch *typische* Information („per Default“) instanziiert werden können. Solche Defaults können durch aktuell vorliegende Information überschrieben werden, charakterisieren aber insgesamt den *Prototyp* eines Frames. Frames dienen einerseits der Klassifikation vorliegender Information und andererseits dem Inferieren weiterer Information: da sie in hierarchischer Beziehung zueinander stehen (und somit *Taxonomien* darstellen), kann Information an untergeordnete Frames **„vererbt“** werden.

Das **Logik-Paradigma** betont die Uniformität der Darstellung von Wissen (in erster Linie mit Hilfe der Prädikatenlogik erster Stufe oder Varianten davon), insbesondere auch für die Anwendbarkeit allgemeiner Inferenzmechanismen (Theorembeweiser). Sein Nachteil besteht in der Strukturarmut logischer Repräsentationen.

Das **(Produktions-)Regel-Paradigma** fokussiert auf den Aspekt der Steuerung des Verhaltens eines Systems durch die Anwendung von (Wenn-Dann-)

Regeln (Inferenzregeln) auf jeweils aktuelle Daten.

Moderne Wissensrepräsentationssysteme wie z. B. PowerLoom® (`http://www.isi.edu/isd/LOOM/PowerLoom/index.html`) stellen oft eine Mischung aus diesen Paradigmen dar.

1.4.3 Struktur und Aufbau von Wissensbasen

Insbesondere den frühen semantischen Netzwerken mangelte es generell an formaler Klarheit der in ihnen verwendeten Repräsentationskonstrukte, vor allem der Kanten (vgl. den Titel von Woods 1975, „What's in a link"). Auch wenn der Einsatz der Netzwerke teilweise zu beeindruckenden Ergebnissen führte (wie die Verwendung der *konzeptuellen Dependenzstrukturen* Roger Schanks für das Textverstehen, s. Schank 1975), blieb ein wesentlicher Teil ihrer Bedeutung oft in ihrem Verarbeitungsmechanismus versteckt.

Ein abschreckendes Beispiel dafür ist das in Abb. 1 dargestellte Netzwerk. Hier stellt sich die Frage, welches Wissen repräsentiert sein soll: Dass alle Menschen, die Peter heißen, Popcorn essen? Oder handelt es sich um einen bestimmten Peter? Ist Popcorn notwendigerweise etwas Essbares oder so zufällig wie das Menschsein von Peter (der ja auch ein Hund sein könnte)? Findet die Handlung immer im Kino statt oder handelt es sich um eine spezifische Situation? Von welchen Kanten kann noch eine loc-Kante ausgehen?

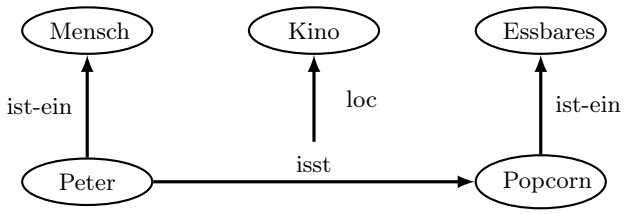

Abbildung 1: Beispiel für ein inadäquates semantisches Netzwerk

Nach Brachman (1979) muss stattdessen systematisch zwischen den inhaltlichen und strukturellen Aspekten von Wissensrepräsentationen unterschieden werden: Während der *Inhalt* nicht-sprachlichen Wissens auf der ***konzeptuellen* Ebene** eines Wissensrepräsentationssystems spezifiziert wird, wird dessen *Form* bzw. *Struktur* auf der ***epistemologischen* Ebene** determiniert.

Die epistemologische Ebene adressiert somit das strukturelle Inventar von Netzwerken und deren Wohlgeformtheit. Entsprechend werden hier die Kanten definiert, die die hierarchischen Beziehungen zwischen Konzepten (Subkonzept-Superkonzept, Individuenkonzept-Klassenkonzept) oder deren Eigenschaften (die „Rollen" von Konzepten (Relationen, Attribute)), darstellen. Als Resultat dieser Überlegungen ergibt sich, dass einige der Kanten „objektifiziert" werden: z. B. vermitteln Ereigniskonzepte, denen räumliche/zeitliche Information attribuierbar sind, zwischen Nominalkonzepten.

Dieser grundlegende Unterschied zu simplen Netzwerken ist in Abb. 2 auf der nächsten Seite anhand eines KL-ONE-ähnlichen Netzwerks (s. Brachman and

Schmolze 1985) veranschaulicht. Sie zeigt, dass sich selbst Rollen als Objekte auffassen lassen, die über rein strukturelle Kanten mit Konzepten verbunden sind. Dies hat zwei Vorteile: Erstens können Rollen selbst Eigenschaften haben, z. B. „v/r (value restriction)“ als Beschreibung ihres Wertebereichs, den typischen Wert oder ihre Häufigkeit. Zweitens können sie, wie die Konzepte, hierarchisch organisiert sein.

In Bezug auf das konzeptuelle Wissen muss nach Helbig (2001) zwischen dem, was es gibt und was immer wahr ist (*immanentes Wissen*), und dem, was (nur) in einer bestimmten Situation der Fall ist (*situatives Wissen*), unterschieden werden. Das immanente Wissen umfasst sowohl die notwendigen Merkmale eines Konzepts (z. B. dass Popcorn etwas Essbares ist) wie auch die das *Defaultwissen* ausmachenden (z. B. dass Menschen im Kino Popcorn essen).

Gleichzeitig muss es möglich sein, Strukturen zu bilden, die den Defaultannahmen widersprechen, u. a. um spezifische Aussagen (*Assertionen*) zu ermöglichen. Dies ist in Abb. 2 am Beispiel des Satzes „Peter isst im Cinestar Erdnüsse“ dargestellt. Man beachte, dass hier jedoch nicht alle relevanten Aspekte expliziert sind.

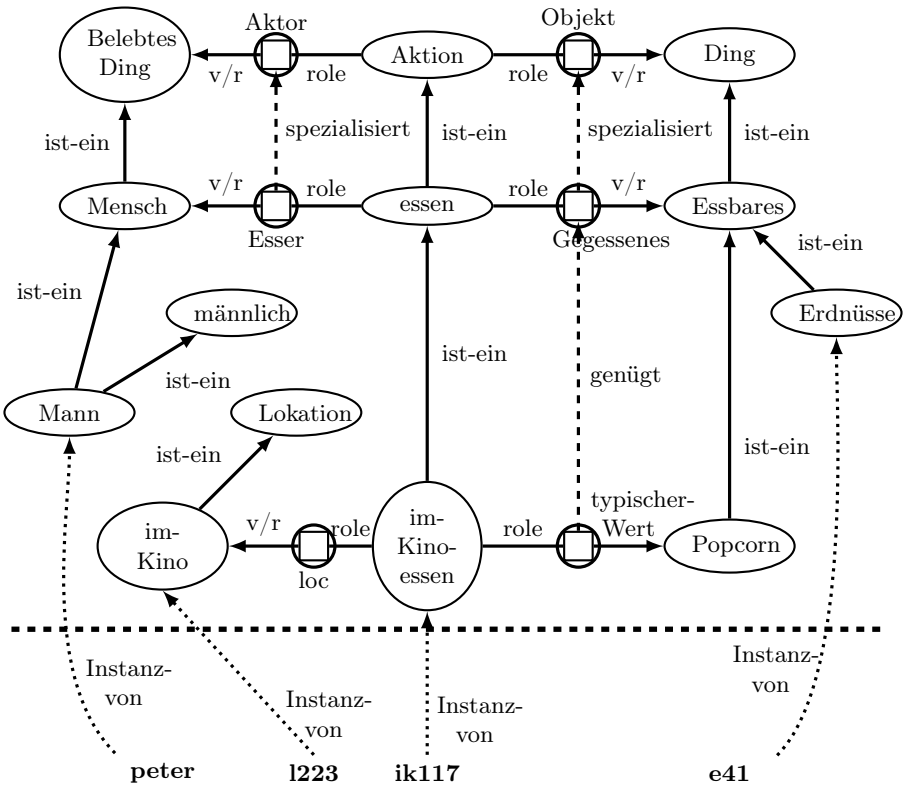

Abbildung 2: Elaborierte Netz-Darstellung repräsentierten Wissens

Hieraus resultiert die Unterscheidung einer terminologischen Komponente („**T-Box**“) und einer assertionalen, Faktenwissen repräsentierenden Komponente („**A-Box**“) in Frame-basierten Wissensrepräsentationssystemen wie KL-ONE (Brachman and Schmolze 1985)—eine Zweiteilung, die in Abb. 2 durch eine waagerechte, gestrichelte Linie dargestellt ist. Die T-Box enthält danach die Konzeptdefinitionen, die A-Box die mithilfe dieser Konzepte gebildeten Aussagen über Individuen einer gegebenen Domäne. Aus dieser Konstellation ergeben sich Informationen über Individuen als Instanzen bestimmter Konzepte sowie (hier nicht dargestellt) von Paaren als Instanzen bestimmter Rollen.

Üblicherweise werden T-Box und A-Box, ggf. zusammen mit einer Regelkomponente, als die **Wissensbasis** eines entsprechenden Systems bezeichnet. **Wissensbasierte Systeme** verfügen neben der Wissensbasis über einen **Inferenzmechanismus**, der die Wissensbasis manipulieren kann (zum Aufbau einer Wissensbasis s. Brachman et al. 1990). Natürlichsprachliche Systeme verfügen zudem über ein **Lexikon** (mit Verweisen sprachlicher Ausdrücke auf Konzepte in der T-Box) und ein **Onomastikon** (mit Verweisen von Namen auf bestimmte Instanzen in der A-Box).

Dem Desiderat formaler Klarheit folgend gilt es seither als Maxime, Sprachen zur *Beschreibung* von Repräsentationskonstrukten mit einer entsprechend expliziten Semantik zu versehen, die einen Bezug der Konstrukte zu ihrer Interpretation in der Welt herstellt (z. B. die Interpretation von Konzepten als Mengen von Individuen und von Rollen als Mengen von Paaren). Heraus resultiert der Terminus **Beschreibungslogik** (*description logic*) für entsprechende formale Sprachen zur Spezifizierung von Wissensrepräsentationsformalismen.

Description logics (DL) sind –üblicherweise stark restringierte– Versionen der Prädikatenlogik, mithilfe derer sich Wissensrepräsentationskonstrukte formal darstellen lassen. Abb. 3 zeigt einen Ausschnitt einer DL-Formalisierung von Abb. 2.

$$
\begin{array}{|l|}
\hline
\mathbf{T-Box}: \\
Mann \doteq Männlich \sqcap Mensch \\
essen \doteq Aktion \sqcap \forall Gegessenes.Essbares \sqcap \forall Esser.Mensch \\
Popcorn \sqsubset Essbares \\
\ldots \\
- \\
\mathbf{A-Box}: \\
Mann(peter) \\
im-Kino-essen(ik117) \\
Esser(ik117, peter) \\
Gegessenes(ik117, e41) \\
Erdnüsse(e41) \\
loc(ik117, l223) \\
\ldots \\
\hline
\end{array}
$$

Abbildung 3: T-Box und A-Box- Ausschnitte zu Abb. 2

Konzepte können über eine Konjunktion von Konzeptbeschreibungen *definiert* werden, wobei Konzeptbeschreibungen u. A. Konzeptnamen (wie bei der Definition von $Mann$) oder Rollenbeschreibungen (wie bei der Definition von *essen*) sein können. Die angegebenen Rollenbeschreibungen formalisieren die Wertebeschränkungen für die jeweilige Rolle. So ist *essen* definiert als eine Aktion, für deren Rolle *Gegessenes* alle Werte Instanzen von $Essbares$ und für deren Rolle *Esser* alle Werte Instanzen von $Mensch$ sind.

Werden Konzepte nicht definiert, so lassen sich wie bei *Popcorn Subsumptionsbeziehungen* angeben (alle Instanzen von *Popcorn* sind Instanzen von $Essbares$). Subsumptionsbeziehungen bilden das Gerüst hierarchisch organisierter Netzwerke. Daher wird beim Aufbau einer Wissensbasis durch einen Mechanismus (sog. „**classifier**“) systematisch gecheckt, ob diese Beziehungen gelten und wo ggf. ein neues Konzept in der Hierarchie angesiedelt ist. Konzept*definition* entspricht im Übrigen einer gegenseitigen Subsumption von Definiens und Definiendum.

Eine dritte Möglichkeit, Konzepte einzuführen, besteht darin, sie nicht weiter zu beschreiben, sondern sie als gegeben/primitiv zu deklarieren. In einer ontologischen Struktur führt das dazu, dass sie direkt unter dem allgemeinsten Konzeptknoten angesiedelt sind.

1.4.4 Ontologien

Die Konstruktion von Wissensbasen erweist sich als ein komplexes Unterfangen, das prinzipiell einen großen Spielraum für Beliebigkeit und nachfolgende Inkonsistenzen lässt. Gerade unter dem Aspekt der Wiederverwendbarkeit von Wissensbasen ist es daher wichtig, Einheitlichkeit und Konsistenz zu gewährleisten.

Die Lösung besteht darin, sich auf eine gemeinsame Sicht der Welt (sog. „Konzeptualisierung“) zu einigen und ebendiese zu formalisieren, wie es seit langem in der philosophischen Disziplin der *Ontologie* versucht wird: „Ontology as a branch of philosophy is the science of what is, of the kinds and structures of objects, properties, events, processes and relations in every area of reality.“ (Smith 2003, S. 155).

In der Wissensrepräsentation wird daran angelehnt unter einer **Ontologie** ein aus solchen Überlegungen resultierendes System von Konzepten (Begriffen) verstanden, das unser Wissen über die (Struktur der) Welt enthält (dabei werden auch Fragen der *kognitiven* Repräsentation von Wissen relevant, s. dazu z. B. Carstensen 2007, Evermann 2005 und Markman 1999).

Diese in der CL, KI und verwandten Disziplinen mittlerweile weit verbreitete Auffassung von „Ontologie“ wurde in Gruber (1995) durch die Formulierung „*explizite Spezifikation einer gemeinsamen Konzeptualisierung*“ geprägt, d.h. als eine formale, sprachunabhängige Beschreibung einer intersubjektiven Sicht der Welt, die als gemeinsamer Kern verschiedener Wissensbasen dienen kann (s. aber Smith 2003 zu Unterschieden dieser und einer realistisch-philosophischen Auffassung von „Ontologie“).

Eine Ontologie, die den Oberbau von T-Boxen darstellt (daher auch: *upper-level* oder *top ontology*) beschränkt sowohl die Primitive der konzeptuellen Ebene bzgl. ihrer möglichen Interpretationen als auch die möglichen Optionen auf der epistemologischen Ebene. Kriterien für deren Wohlgeformtheit lassen sich nach Guarino (1995) aus philosophischen und linguistischen Überlegungen gewinnen. Zwischen diesen beiden Ebenen ist deshalb nach Guarino eine eigenständige ***ontologische* Ebene** anzunehmen.

Mittlerweile versteht man unter einer Ontologie nicht nur diesen Kern, sondern auch die übrigen Wissensrepräsentationsstrukturen, also auch bereichsspezifische (*domain ontologies*), zwischen upper-level und domain-level vermittelnde (*mid-level ontologies*) oder aufgabenspezifische (*task ontologies*) (s. hierzu auch die EuroWordNet-Architektur in Unterkapitel **??**). In jedem Fall kommt aber der upper/mid-level ontology als wiederverwendbarer Ressource eine besondere Bedeutung zu (Beispiele sind die Suggested Upper Merged Ontology SUMO und die MId-Level Ontology MILO, s. hierzu `http://www.ontologyportal.org`).

In einem noch weiter gefassten Sinn werden zum Teil alle Wissensressourcen eines wissensbasierten Systems —beispielsweise lexikalisch-semantische Ressourcen wie WordNet (s. Unterkapitel **??**)— als Ontologien bezeichnet.

1.4.5 Von der Wissensrepräsentation zum Semantischen Web

Nach dem Aufkommen der semantischen Netzwerke (Ende der 60er Jahre) und des Frame-Paradigmas (Anfang der 70er Jahre) setzte eine „Logifizierung“ der Wissenrepräsentationsformalismen ein, aus der verschiedene (Default-)Logiken sowie die Beschreibungslogiken (auch: *terminologische Logiken*) resultieren. Die deutlich werdende „Zerbrechlichkeit“ der wissensbasierten Systeme (vor allem der Expertensysteme) führte 1984 zu dem Projekt Cyc (von „Encyclopedia“, s. Lenat and Guha 1989), in dem, ausgehend von der Entwicklung einer top ontology, eine umfassende, „common sense knowledge“ repräsentierende Wissensbasis entwickelt werden sollte (und immer noch wird). Die 90er Jahre sahen ein rasant ansteigendes Interesse an Ontologien, ihrer Erstellung und Verarbeitung (sog. „ontological engineering“), insbesondere auch unter dem Aspekt der Wiederverwendbarkeit von Wissen im oder für das WWW.

Berners-Lee et al. (2001) veranschaulichen das Problem für Computer, die Daten einer Webseite zu verstehen und entsprechend zu bearbeiten. Sie entwickeln die Vision eines **Semantic Web**, in dem die Daten des Web mit Information über ihre jeweilige Bedeutung annotiert sind. Mittlerweile existieren hierfür Vorschläge, die auf einer systematischen Schichtung (sog. „Semantic layer cake“) immer mächtigerer Beschreibungssprachen basieren. Etwa in der Mitte zwischen Webdaten-Implementation und spezifischen Anwendungen ist die Web Ontology Language (OWL, s. McGuinness and van Harmelen 2004) angesiedelt, mit der sich (onto)logische Ausdrücke der description logic darstellen lassen. Sie basiert auf dem Resource Description Format (RDF), mithilfe dessen Wissensrepräsentationskonstrukte (Konzepte, Rollen, Instanzen) als Webdaten definiert und schließlich im XML-Format für Annotierungen zur Verfügung gestellt wer-

den können.

Eines der drängenden Probleme bei der Entwicklung des Semantic Web ist die automatisierte Erstellung von Ontologien. Zu dessen Lösung etabliert sich ein zwischen Information Retrieval, Künstlicher Intelligenz und Computerlinguistik angesiedelter Bereich namens **ontology learning and population from text**, in dem ontologisches Wissen aus den Texten von Webseiten induziert wird (s. Cimiano 2006 sowie Unterkap. **??** zum WWW als Ressource).

Als separate Entwicklung lässt sich auch eine Renaissance früher Wissensrepräsentationsideen für die Klassifikation und das selektive Retrieval von Web-Dokumenten unter dem Stichwort **Topic maps** beobachten: Topic maps sind das Analogon der frühen semantischen Netzwerke im Bereich des WWW. Sie bestehen aus einer abstrakten Ebene von *topics* (Themen, von denen Webseiten handeln können), die durch *associations* miteinander verknüpft sind. Insbesondere das Konzept der associations weckt allerdings die Erinnerung an die klassischen Probleme der Wissensrepräsentation (derer sich seine Erfinder möglicherweise nicht bewusst sind), weswegen Topic Maps kaum eine ernsthafte Rolle in zukünftigen Versionen des Web spielen werden.

1.5 Wissensrepräsentation für die CL

1.5.1 Probleme

Die Verwendung nicht-sprachlichen Wissens in natürlich-sprachlichen Systemen ist grundsätzlich nicht-trivial, da die existierenden Ressourcen weder ausgereift noch generell mit sprachlichen Komponenten kompatibel (d.h. für natürlich-sprachliche Zwecke geeignet) sind. Im Einzelnen lassen sich die folgenden Probleme nennen:

Die notwendige **Trennung nicht-sprachlicher und sprachlicher Konzepte** wird nicht immer strikt eingehalten. Zum Einen werden in der Wissensmodellierung sprechende Bezeichner (d.h. sprachliche Symbole) verwendet, was zumindest potentiell (wohl aber sogar faktisch) zu einem sprachlichen und auch kulturspezifischen „Bias“ führt (wodurch insbesondere bei Interlingua-basierter Übersetzung Probleme auftreten können).

Zum Anderen werden sprachliche Ressourcen wie WordNet nicht selten als Ontologien im engeren Sinne verwendet. Durch Anwendung ihrer OntoClean-Methodologie zeigen Gangemi et al. (2003), dass WordNet für diesen Zweck nicht geeignet ist, da dessen Konzeptstruktur einigen grundlegenden ontologischen Wohlgeformtheitsbedingungen nicht genügt.

Dem Leser wird nicht entgangen sein, dass diese Aspekte nur Spezialfälle der allgemeinen terminologischen Verwirrung in der Redeweise darüber sind, wie die Welt (→ ontologisch), unser Wissen über die Welt (→ konzeptuell, epistemologisch) und die sprachlichen Ausdrücke (→ terminologisch) beschaffen bzw. zu modellieren sind.

Die **Nicht-Berücksichtigung sprachlich relevanter Differenzierungen** in Ontologien bedeutet deren Unbrauchbarkeit für computerlinguistische Zwecke. Entsprechend müssen ausschließlich von spezifischem Domänenwissen

abstrahierende Top-Ontologien um sprachlich relevante Konzepte erweitert werden. Hierzu gehört z. B. die systematische Unterscheidung (quantifizierbarer) Massen von (zählbaren) Objekten, um die Inadäquatheit von **Peter isst viel Apfel.* und **Viel Menschen isst einen Popcorn.* erklären und systematisch ausschließen zu können. Ein Vorschlag für eine entsprechende, an linguistischen Phänomenen orientierte upper-level Ontologie ist das *Generalized Upper Model* von Bateman et al. (1994).

Grundlegende Probleme der Wissensmodellierung, wie z. B. die **Bestimmung eines einheitlichen Inventars an Repräsentationsprimitiven**, sind immer noch nicht gelöst. So zeigt ein Vergleich von Wissensrepräsentationssystemen eine erschreckende Uneinheitlichkeit schon bei grundlegenden Repräsentationskonstrukten.

Ein besonderes Problem ergibt sich aus der weitgehenden **Arbitrarität konzeptueller Slots** im Frame-Paradigma. Ein Beispiel dafür findet sich im System **Cyc**: Es enthält tatsächlich Wissen darüber, wie man Popcorn isst, und zwar in Form eines entsprechenden ‚`EatingPopcorn`'-Frames. Hierzu gehört nach Lenat and Guha (1989, S. 192) ein Slot ‚`bodyPartsRequiredOfPerformer`', für den die Werte ‚`Teeth, Mouth, Throat, Stomach, Brain`'[!] angegeben sind.

Slots wie diese verstecken komplexe Strukturzusammenhänge hinter sprachlich eingängigen Bezeichnungen. Deren Beziehung zum restlichen Wissen (bzw. deren Semantik) ist entweder nicht vorhanden oder nur durch ein komplexes „Slot-Bookkeeping" herzustellen. Solche Probleme haben übrigens zur Aufgabe der Frame-Basiertheit in Cyc zugunsten einer logisch verteilten Repräsentation (sog. „knowledge soup") geführt. Wie Mahesh et al. (1996) in ihrer Analyse der NLP-Tauglichkeit von Cyc zeigen, hat jedoch die daraus resultierende Unstrukturiertheit der Repräsentation ebenfalls negative Konsequenzen für deren Verwendbarkeit in natürlich-sprachlichen Systemen.

Ein immer noch bestehendes praktisches Problem ist die **mangelhafte Verfügbarkeit** (vor allem kommerziell entwickelter) vorhandener Ressourcen. Ausnahmen sind beispielsweise die Suggested Upper Merged Ontology (SUMO, s. `http://www.ontologyportal.org/`) und der frei verfügbare Teil der Ontologie aus Cyc (OpenCyc, s. `http://www.opencyc.com/`)

1.5.2 Perspektiven

Die Nutzung nicht-sprachlichen Wissens als computerlinguistische Resssource wird, nicht zuletzt durch den bislang ausgebliebenen Erfolg von Cyc, in der nächsten Zeit voraussichtlich vor allem pragmatisch ausgerichtet sein. Dies bedeutet in erster Linie, dass die (schnelle) Erfüllung einiger Erwartungen aufgegeben werden muss, nämlich die Erwartung einer *perfekten* Gesamt-Ontologie, einer *idealen* Interlingua, einer *alles abdeckenden* Top-level Ontologie oder einer *allumfassenden* Wissensbasis. Stattdessen werden existierende Ressourcen trotz partieller Inkompatibilitäten zusammengefügt („ontology merging") wie z. B. bei der Entwicklung der *Sensus*-Ontologie (Knight and Luk 1994).

Aufgrund der festgestellten Lücken und Qualitätsunterschiede in der Wissensbasis von Cyc schlägt Mahesh (Mahesh and Nirenburg 1995; Mahesh 1996)

im Rahmen der auf maschinelle Übersetzung ausgerichteten Entwicklung der *Mikrokosmos*-Ontologie einen *situierten* Ansatz der Ontologie-Konstruktion vor, bei dem die Wissensmodellierung insbesondere von der Aufgabe und dem frühen praktischen Einsatz der Wissensbasis geleitet wird. Dabei werden sprachlich relevante Unterscheidungen einerseits von vornherein und andererseits so weit wie nötig bei der Ontologie-Konstruktion berücksichtigt, so dass die Erstellung einer ontologischen Interlingua einen *approximativen* Charakter erhält. Im Gegensatz zur manuellen Erstellung von Ontologien wird die Bildung/Population von Ontologien mittlerweile allerdings automatisch anhand von aus Texten gewonnenen Informationen vorgenommen („ontology learning“, s. Cimiano 2006).

Komplementär dazu verfolgt Guarino eine formal-ontologisch gesteuerte Wissensmodellierung (Guarino 1995), indem er ontologische Meta-Eigenschaften als Constraints für wohlgeformte Ontologien verwendet. Zusammen mit den intensivierten informatischen Bestrebungen auf den Gebieten Standardisierung von Wissensrepräsentationsformalismen und Tool-Development (Ontologie-Browser etc.) lassen diese Entwicklungen erhebliche Qualitätsverbesserungen erwarten.

Angesichts der Tatsache, dass auch der Ansatz Guarinos stark von linguistischen Überlegungen geprägt und dass das ontology learning im Kern ein sprachtechnologischer Zweig ist, sind es paradoxerweise die (Computer-)Linguisten, die Wesentliches im Bereich nicht-sprachlichen Wissens beitragen.

1.6 Literaturhinweise

Standardwerke der Einführung in die Wissensrepräsentation im Hinblick auf die Verarbeitung natürlicher Sprache sind Sowa (1984) und Sowa (2000). Eine lesbare Einführung mit vielen Beispielen ist Reimer (1991). Empfehlenswert im Hinblick auf Aktualität, Umfang und Detailreichtum sprachlich orientierter Wissensrepräsentation ist Helbig (2001). Interessierte finden die klassischen Papiere zur Wissensrepräsentation in Brachman and Levesque (1985) und eine umfassende Einführung in die Wissensrepäsentation und -verarbeitung aus Sicht der Künstlichen Intelligenz in Brachman and Levesque (2004). Einen Einstieg in das Thema „description logics“ liefert Baader et al. (2003) sowie die Webseite `http://dl.kr.org`.

John Bateman (`http://www-user.uni-bremen.de/~bateman/`) bietet eine Sammlung relevanter Information zu Ontologien und weiteren Aspekten der Wissensrepräsentation an, die insbesondere im Hinblick auf die Entwicklung von Grundlagen für das Semantic Web relevant ist (s. dazu Hitzler et al. 2007, `http://www.w3.org/RDF/FAQ` sowie auch Unterkapitel **??**). Verschiedenste Perspektiven auf den Bereich Semantic Web (und Ontologien) finden sich in Pellegrini and Blumauer (2006). Aus (computer)linguistischer Perspektive besonders interessant ist die Ontologie DOLCE („Descriptive Ontology for Linguistic and Cognitive Engineering“, s. `http://www.loa-cnr.it/DOLCE.html`).

Die praktische computerlinguistische Verwendung solcher Ontologien wird z. B. in Oberle et al. (2007) präsentiert. Nirenburg and Raskin (2004) enthält eine umfassende computerlinguistische Darstellung der Rolle von Ontologien für die Computerlinguistik und Sprachtechnologie.

Literatur

Baader, F., D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider (2003). *The Description Logic Handbook: Theory, Implementation, Applications*. Cambridge University Press, Cambridge, UK.

Bateman, J. A., B. Magnini, and F. Rinaldi (1994). The generalized italian,german,english upper model. In *Proceedings of the ECAI94 Workshop: Comparison of Implemented Ontologies*, Amsterdam.

Berners-Lee, T., J. Hendler, and O. Lassila (2001, May). The semantic web. *Scientific American 284*(5), 24–30.

Brachman, R. and J. Schmolze (1985). An overview of the KL-ONE knowledge representation system. *Cognitive Science 9*, 171 – 216.

Brachman, R. J. (1979). On the epistemological status of semantic networks. In N. Findler (Ed.), *Associative Networks*, pp. 3–50. New York: Academic Press.

Brachman, R. J. and H. J. Levesque (1985). *Readings in Knowledge Representation*. Los Altos, CA: Morgan Kaufmann.

Brachman, R. J. and H. J. Levesque (2004). *Knowledge Representation and Reasoning*. Morgan Kaufmann.

Brachman, R. J., D. L. McGuiness, P. F. Patel-Schneider, and L. A. Resnick (1990). Living with CLASSIC: when and how to use a KL-ONE-like language. In J. Sowa (Ed.), *Principles of semantic networks*. San Mateo, US: Morgan Kaufmann.

Carstensen, K.-U. (2007). Spatio-temporal ontologies and attention. *Spatial Cognition & Computation 7*(1), 13–32.

Cimiano, P. (2006). *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer Verlag.

Evermann, J. (2005). Towards a cognitive foundation for knowledge representation. *Information Systems Journal 15*(2), 147–178.

Gangemi, A., N. Guarino, C. Masolo, and A. Oltramari (2003). Sweetening WordNet with DOLCE. *AI Magazine 24*(3), 13–24.

Gruber, T. R. (1995). Towards principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies 43*, 907 – 928.

Guarino, N. (1995). Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human and Computer Studies 43*(5-6), 625–640.

Harnad, S. (1990). The symbol grounding problem. *Physica D 42*, 335 – 346.

Helbig, H. (2001). *Die semantische Struktur natürlicher Sprache*. Springer-Verlag.

Hitzler, P., M. Krötzsch, S. Rudolph, and Y. Sure (2007). *Semantic Web: Grundlagen*. Springer.

Hitzler, P. and K.-U. Kühnberger (2009). The importance of being neural-symbolic – a wilde position. In B. Goertzel, P. Hitzler, and M. Hutter (Eds.), *Artificial General Intelligence*, Second Conference on Artificial General Intelligence, AGI, Arlington, Virginia, USA.

Knight, K. and S. Luk (1994). Building a large knowledge base for machine translation. In *Proceedings of American Association of Artificial Intelligence Conference (AAAI-94)*, Seattle, WA, pp. 773–778.

Lang, E., K.-U. Carstensen, and G. Simmons (1991). *Modelling Spatial Knowledge on a Linguistic Basis*. Springer-Verlag. Lecture Notes in Artificial Intelligencce No. 481.

Lenat, D. and R. Guha (1989). *Building large knowledge-based systems. Representation and inference in the Cyc project*. Reading, Mass.: Addison-Wesley.

Mahesh, K. (1996). Ontology development for machine translation: Ideology and methodology. Technical Report MCCS-96-292, NMSU CRL.

Mahesh, K. and S. Nirenburg (1995). A situated ontology for practical NLP. In *Proc. Workshop on Basic Ontological Issues in Knowledge Sharing,International Joint Conference on Artificial Intelligence (IJCAI-95)*, Montreal, Canada.

Mahesh, K., S. Nirenburg, J. Cowie, and D. Farwell (1996). An assessment of Cyc for natural language processing. Technical Report MCCS-96-302, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.

Markman, A. B. (1999). *Knowledge representation*. Mahwah, NJ: Lawrence Erlbaum Associates.

McGuinness, D. and F. van Harmelen (2004). OWL web ontology language overview. Technical report, W3C Recommendation.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The Psychology of Computer Vision*, pp. 211–277. New York: McGraw Hill.

Newell, A. and H. Simon (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM 19*, 113 – 126.

Nirenburg, S. and V. Raskin (2004). *Ontological Semantics*. The MIT Press.

Oberle, D., A. Ankolekar, P. Hitzler, P. Cimiano, M. Sintek, M. Kiesel, B. Mougouie, S. Vembu, S. Baumann, M. Romanelli, P. Buitelaar, R. Engel, D. Sonntag, N. Reithinger, B. Loos, R. Porzel, H.-P. Zorn, V. Micelli, C. Schmidt, M. Weiten, F. Burkhardt, and J. Zhou (2007). DOLCE ergo SUMO: On foundational and domain models in the smartweb integrated ontology (SWIntO). *Journal of Web Semantics: Science, Services and Agents on the World Wide Web 5*(3), 156–174.

Pellegrini, T. and A. Blumauer (Eds.) (2006). *Semantic Web: Wege zur vernetzten Wissensgesellschaft*. Berlin: Springer Verlag.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic Information Processing*, pp. 227–270. Mass.: MIT Press.

Reimer, U. (1991). *Einführung in die Wissensrepräsentation*. Stuttgart: B.G.Teubner.

Schank, R. C. (1975). *Conceptual Information Processing*. Amsterdam: North-Holland.

Smith, B. (2003). Ontology. In L. Floridi (Ed.), *Blackwell Guide to the Philosophy of Computing and Information*, pp. 155–166. Oxford: Blackwell.

Smith, B. C. (1982). Prologue to 'reflection and semantics in a procedural language' [reprinted in Brachman and Levesque (1985)]. Technical Report 272, MIT.

Sowa, J. F. (1984). *Conceptual Structures : Information Processing in Mind and Machine*. Reading, Mass.: Addison-Wesley.

Sowa, J. F. (2000). *Knowledge representation: logical, philosophical, and computational foundations*. Pacific Grove, CA.: Brooks/Cole.

Woods, W. A. (1975). What's in a link: foundations for semantic networks. In D. Bobrow and A. Collins (Eds.), *Representation and Understanding*, pp. 35–82. New York: Academic Press.