

## Kompositionalität von Begriffen – Bedeutung und formale Aspekte

Straub HR<sup>1</sup>, Badertscher M<sup>2</sup>, Frei N<sup>2</sup>

<sup>1</sup>Semfinder AG, Kreuzlingen, Schweiz

<sup>2</sup>Interstaatliche Hochschule für Technik, Buchs, Schweiz

straub@semfinder.ch

**Einleitung** Ein Begriff, z.B. eine Diagnose, setzt sich aus mehreren Begriffen zusammen, so ist z.B. eine *Zystitis* eine *Entzündung der Harnblase*; die beiden letztgenannten Begriffe sind im ersten enthalten. Kompositionalität bedeutet dabei zweierlei: a) Welche Teilbegriffe sind vorhanden (Aufzählung) und b) Wie sind sie angeordnet (Strukturierung). Durch den Zugriff auf die Teilkonzepte können Abfragen einfach und systematisch erfolgen und durch die Strukturierung der Begriffsverhältnisse können auch komplexe Aufgaben rechnergestützt durchgeführt werden. Die Anordnung der Begriffskomponenten ist dabei die entscheidende Herausforderung. Im Folgenden zeigen wir die Prinzipien der Kompositionalität in den Begriffsmolekülen (BM). Begriffsmoleküle sind insbesondere bezüglich Lesbarkeit und Wartbarkeit von sehr grossen und komplexen Wissensbasen optimiert [9].

**Begriffe und Instanzen** Hier herrscht eine gewisse Verwirrung; Begriffe und Instanzen müssen mit allen Konsequenzen klar unterschieden werden. Die Kompositionalität der Begriffe ist nicht gleichbedeutend mit der Kompositionalität von Instanzen. So enthält der Begriff *Zystitis* den Begriff *Harnblase*, die *Zystitis* als reale Instanz aber keine *Harnblase*, sondern sie betrifft eine solche. Die Anordnung der Begriff darf nicht mit der Anordnung der Objekte verwechselt werden, das Part-Of der Begriffe bedeutet nicht das Part-Of der Instanzen, und umgekehrt.

**Begriffshierarchien, Mehrdimensionalität** Für Auswertungszwecke und für die maschinelle Verarbeitung ist eine hierarchische Anordnung ideal, allerdings finden sich hierarchische Kategorienbäume in der Realität nicht so problemlos [7] wie in der Theorie [4,8]. Trotzdem sind hierarchische Ketten sinnvoll und bilden die Basisstruktur der Begriffsmoleküle (BM). Komplexe Wissensgebiete lassen sich durch die Verwendung von *mehreren* Hierarchien (=Dimensionen) besser abbilden, bei medizinischen Diagnosen sind diese z.B. die Organe, Krankheitsursachen, Lokalisationen, Progressionstypen usw. Aus jeder Dimension kann für die Spezifizierung einer konkreten Diagnose jeweils ein unabhängiger Wert verwendet werden (komposite Terminologie).

**Bifazialität** Ein Unterbegriff steht zu einem Oberbegriff in einer Is-A Relation. Wenn in einer hierarchischen Kette drei und mehr Begriffe aufeinander folgen, haben die Zwischenbegriffe beide Funktionen, sie sind gleichzeitig (für den spezifischeren Unterbegriff) Oberbegriff und (für den allgemeineren Oberbegriff) Unterbegriff. Beispiel einer solchen Kette ist: *Diagnose* → *Neoplasie* → *Malignom* → *Karzinom* → *Lungenkarzinom*. Die Doppelfunktion der Zwischenbegriffe muss formal klar dargestellt werden können. Bei OWL DL muss für jedes Element angegeben werden, welche Funktion auf das Element zutrifft, ob es eine Klasse ist oder ein Individual [1,2,5]. Bei Begriffsmolekülen (BM) [9] und OWL Full [2] kann jedes Atom beide Funktionen tragen, alle Werte (Begriffe) sind potentiell sowohl Klasse wie Individuale. Bei Begriffsmolekülen gilt dies im Gegensatz zu OWL Full sehr streng: jedes Atom ist prinzipiell sowohl Klasse wie Individual. Formal werden die beiden Funktionen als zwei obligat angelegte und eindeutig unterscheidbare Bindungsstellen in jedem atomaren Begriffselement repräsentiert (Optimierung von Expressivität und Maschinengängigkeit). In der optischen Darstellung der BM zeigt die Bindung für Oberbegriffe obligat nach links, diejenige für Unterbegriffe nach rechts. Hierarchische Ketten von beliebiger Länge sind möglich. Wenn ein Begriff als Individual in die Repräsentation aufgenommen wird (*Influenza*), ist es später möglich, den Begriff als Klasse zu verwenden und mit neuen Unterbegriffen zu spezifizieren (*Vogelgrippe*), ebenso ist es möglich, die hierarchische Kette in der Mitte aufzubrechen und Zwischenbegriffe einzufügen.

**Ahnenpersistenz** Begriffe zeigen eine verblüffende Verwandtschaft mit den Objekttypen der objektorientierten Programmierung (OOP) [3]. Der Unterbegriff entspricht dem abgeleiteten Objekttyp (*Sohn*) der OOP, der Oberbegriff dem *Vatertyp*. Bei der OOP erbt der *Sohntyp* alle Eigenschaften des *Vaters* und in einer Realisierung (Instanz) des Objekttyps wird allein der direkte Typ, d.h. der *Sohntyp* angegeben. Begriffsmoleküle unterscheiden sich hier von den Darstellungen der OOP durch die "Ahnenpersistenz": Bei BM erbt der *Sohn* die Eigenschaften (Variablen) der *Väter* nicht, sondern die gesamte Ahnenkette wird simultan und explizit dargestellt. Dies entspricht der ganz anderen Aufgabe der Begriffsrepräsentation im Vergleich zur OOP und optimiert bei einer Repräsentation von sehr vielen Objekttypen (Begriffen) die Übersicht. Die Ahnenkette wird als einfache Aneinanderreihung der Begriffe in Form der oben erwähnten hierarchischen Kette horizontal am Bildschirm gezeigt (Abb.1).

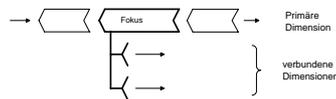


Abb. 1: Verknüpfung von drei Dimensionen über ein Glied (Fokus) der ersten Dimension

**Verknüpfung von Hierarchien** Die oben beschriebene Mehrdimensionalität ist notwendig, aber nicht hinreichend für die Repräsentation von medizinischen Diagnosen. Die Information liegt nicht nur in Begriff (Knoten), bzw. im Bündel der Begriffe (komposite Terminologie, Mehrdimensionalität), sondern auch in den Verknüpfungen, d.h. in den Relation sowie in den Relationen der Relationen. Relationen und Relationen von Relationen bedeuten nichts anderes als die Struktur der Begriffe, d.h. ihre Kompositionalität. Eine nur mehrdimensionale Terminologie, welche die Bezüge zwischen den Dimensionen nicht organisiert, kann die medizinische Begrifflichkeit nicht hinreichend modellieren. In den Begriffsmolekülen (BM) wird die Strukturierung als Verknüpfung von verschiedenen Hierarchien (Dimensionen) explizit dargestellt. Den Punkt, an dem die Verknüpfung stattfindet, nennen wir *Fokus* (Abb.1), das gesamte System ein *multifokales*, im Gegensatz zum nur mehrdimensionalen, das die einzelnen Achsen nicht explizit verknüpft. Die erste Zeile von Abb. 1 zeigt eine hierarchische Kette, deren zweites Glied ein *Fokus* ist, der zu weiteren Dimensionen (Hierarchien) führt. Selbstverständlich ist es möglich, dass auch die anderen Glieder (Atome) der Kette als *Foki* agieren und unabhängig voneinander weitere Dimensionen binden. Durch die oben erwähnte Ahnenpersistenz kann die Bindung von weiteren Dimensionen (Attributen) auf dem jeweils aussagekräftigsten Generationenlevel (Granularitätslevel) erfolgen. Eine explizite Verknüpfung von einer Hierarchie mit einer anderen kann auch in weiteren, bereits untergeordnet verknüpften Hierarchien stattfinden, sodass Hierarchiewechsel kaskadenartig erfolgen (Abb. 2). Ein so verknüpftes Konglomerat von mehreren atomaren Konzepten nennen wir ein *Begriffsmolekül* (Abb. 3). Der Vergleich der Abbildungen 3 und 4 zeigen die fraktale Selbstähnlichkeit von Molekül und Atom.

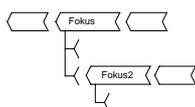


Abb. 2: Verknüpfungen über mehrere Knoten (fraktal)

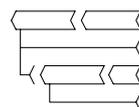


Abb. 3: Molekül

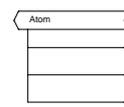


Abb. 4: Atom

**Attributive Slots** In der Notation der BM werden hierarchische Bindungen von Atomen konsequent horizontal von links nach rechts aneinander gereiht (= primäre Dimension in Abb. 1). Unterhalb des Atoms werden die attributiv gebundenen Atome gezeichnet. So besitzt in Abb. 1 das Atom "Fokus" zwei attributive Bindungsstellen (Slots) für zusätzliche, d.h. hierarchiefremde Dimensionen. Jedem Slot ist eine Bindungsliste zugeordnet, welche die Atome der jeweiligen Dimension auflistet. Diese Atome können wieder hierarchisch, d.h. als Baum organisiert werden, wobei das oberste Atom der Hierarchie als Klasse (Bifazialität) das durch den Slot definierte Merkmal quasi benennt (Reification [6]) – so benennen die Atome "Organ" und "Ursache" als Spitze von zwei Hierarchieebenen zwei Slots des Atoms "Diagnose". Jedes Atom besitzt 0-n solche Slots. Die Bedeutung (Semantik) eines Slots definiert sich allein durch die dort gebundenen Atome; dabei ist es egal, ob diese Individuale oder Klassen sind (Bifazialität).

Bei der Wartung der Wissensbasis können Slots wachsen (zusätzliche Bindungen) und die Zahl der Slots kann vermehrt werden. Insgesamt gibt es in der gesamten Wissensbasis nur ein einziges formales Element, nämlich das Atom (Abb. 4).

**Zusammenfassung** Mit BM können vieldimensionale Netze von Begriffen und ihren Verknüpfungen in den zwei Dimensionen von Papier oder Bildschirm dargestellt werden. Der Formalismus kann von Computerprogrammen eindeutig interpretiert werden und erlaubt das Modellieren und Warten von komplexen Wissensbasen. Mit BM kann der komplexe relationale Zusammenhang von 3 und mehr Begriffen – das heisst ihre Kompositionalität einfach und klar dargestellt werden. Die Notation erlaubt insbesondere auch das Erstellen von komplexen Interpretationsregeln, wie sie für die automatisierte Kodierung von Diagnosen und Prozeduren gebraucht werden.

## Literatur

- [1] Baader F et. al (Hrsg). The Description Logic Handbook – Theory, Implementation and Applications. Cambridge: University Press, 2003.
- [2] Dean M, Schreiber G. et. al. OWL Web Ontology Reference – 1.2. OWL Full, DL, Lite. W3C, 2004. <http://www.w3.org/TR/owl-ref/#Sublanguages>
- [3] Dewhurst SC, Stark KS. Programmieren in C++. München, Wien: Hanser, 1990
- [4] Rector AL. Clinical Terminology: Why is it so hard? *Methods Inf Med.* 38(4-5); 1999: 239-52.
- [5] Rector AL. Representing Specified Values in OWL. Value Partitions and Value Sets. 2005. <http://www.w3.org/TR/2005/NOTE-swbp-specified-values-20050517/>
- [6] Rector AL, Rogers J. Ontological & Practical Issues in using a Description Logic to Represent Medical Concepts: Experience from GALEN. The University of Manchester, School of Computer Science: Preprint Series CSPP-35. 2005. see: <http://www.cs.man.ac.uk/cspreprints/PrePrints/cspp35.pdf>
- [7] Smith B. Aristoteles 2002. In: Buchheim T, Hrsg. Kann man heute noch etwas anfangen mit Aristoteles. Hamburg: Meiner, 2003; 3-38.
- [8] Straub HR. Diagnosekodierung als Interpretation sprachlicher Zeichen. In: Schönbächler G: Diagnoseprozesse und Wissenssysteme. *Zeitschrift für Semiotik* 26 (3-4); 2004: 227 –243.
- [9] Straub HR, Frei F, Mosimann H, et. al. Simplified Representation of Concepts and Relations on Screen. In: Proceedings of MIE 2005. Amsterdam: IOS Press; 712-716.