

Lexikalische und ontologische Ressourcen

Claudia Kunze

kunze8@gmail.com, kunze@cl.uni-heidelberg.de

Computational Lexicography

- Typen lexikalischer Datenbanken
- GermaNet als monolinguales lexikalisch-semantisches Wortnetz
- EuroWordNet als polylinguales lexikalisch-semantisches Wortnetz
- FrameNet
- Ontologien

Diese Sitzung befasst sich mit Ontologien.

Warum Ontologien?

- Mensch: kann auf Allgemeinwissen, Fachwissen und Kontextwissen zugreifen, dieses durch Nachschlagen in Lehrbüchern und Lexika ergänzen und mit seinem gespeicherten Wissen verbinden
- Computer: benötigt zur Bearbeitung wissensbasierter Aufgaben, z.B. der Kommunikation von Informationen, Fakten und Sachverhalten, eine formale Struktur, in der die für diese Fragestellung relevanten Begrifflichkeiten und ihre Zusammenhänge verfügbar sind
- Eine solche formale Wissensrepräsentation wird seit Gruber (1993) ONTOLOGIE genannt.

Merkmale einer Ontologie nach Gruber (1993)

Gruber definiert eine Ontologie als „an explicit formal specification of a shared conceptualization of a domain of interest“. Diese Definition enthält vier charakteristische Eigenschaften von Ontologien:

- Darstellung eines konzeptuellen Bereiches, d.h. eines WISSENSBEREICHES in einem standardisierten Vokabular, auch TERMINOLOGIE genannt;
- GEMEINSAME Grundlage für die MODELLIERUNG des Wissensbereiches durch eine Gruppe von Experten;
- Ausrichtung auf eine INTERESSENDOMÄNE, z.B. einen bestimmten Fachbereich oder Anwendungshintergrund;
- explizite FORMALISIERUNG des Wissens, um logisches Schließen über den Daten zu ermöglichen.

Verwendung von Ontologien

- in der Informatik, Künstlichen Intelligenz und Nachbardisziplinen im Bereich der Informationstechnologie;
- in Fachdisziplinen wie der Medizin, Biologie und z.B. der Verwaltung digitaler Bibliotheken;
- in allen Bereichen der Wissensrepräsentation, Wissenskommunikation und des Wissensmanagements inkl. der Behandlung von Metadaten;
- Inferieren über Wissensbestände;
- Wiederverwendung von Wissensbeständen.

Nutzen von Ontologien

- Explizierung von (Domänen)wissen erleichtert den Austausch über die Grundannahmen einer Wissensdomäne; die Begriffsbestimmung über relevante Informationen kann leichter kommuniziert werden.
- Durch Trennung von Domänenwissen und operationalem Wissen ist ein zielgerichteter Zugriff auf diese unterschiedlichen Wissenstypen und deren Wiederverwendbarkeit möglich.
- Mit Ontologien werden Referenzressourcen für Anwendungen in einer Domäne oder für ein spezifisches Anwendungsszenario geschaffen.

Ontologien im SEMANTIC WEB

- besonders populär seit Postulierung des Zeitalters des „Semantic Web“ (vgl. Berners-Lee et al. (2001))
- Auszeichnung von Web-Dokumenten mit semantischen inhaltsbeschreibenden Metadaten und deren Verknüpfung durch Inferenzregeln für intelligente Suchverfahren
- Ontologien können anwendungsspezifische Metadaten und Regeln bereitstellen, z.B. als Grundlage der Kommunikation zwischen Agenten
- Guarino (1998): allgemein ausgerichtete TOP-LEVEL-ONTOLOGIEN vs. spezifische DOMÄNEN-ONTOLOGIEN;
- Begriff der Ontologie umfasst auch die KONZEPTUELLEN GRAPHEN aus der KI (vgl. Sowa (1991) und Brachman (1985)) sowie herkömmliche Daten- und Klassenmodelle.

Annäherung an Ontologien

- historische Grundlage des Ontologiebegriffs und Aristotelische Kategorisierung von Begriffen
- formale Begriffsbestimmung im Rahmen der “Formal Concept Analysis”
- Unterschiede zwischen einer Taxonomie, einem Thesaurus und einer Ontologie
- Aufbau einer Ontologie (exemplarisch)
- Vorstellung einiger für die Sprachwissenschaft relevanter Ontologien

Aristoteles

Aristotelische Metaphysik: ONTOLOGIE als Lehre vom Seienden und der Seinshaftigkeit - Betrachtung der Hintergründe realer Dinge und der abstrakten Entitäten (res abstractae): Ideen, Vorstellungen und Beweggründe.

- Aristoteles als Klassiker der Wissensrepräsentation und des logischen Schließens
- TAXONOMIEN: relevante Begriffe der Ebenenspezifikationen gehen auf Aristoteles zurück:
GENUSTERM als allgemeiner Klassenbegriff zur Kennzeichnung einer natürlichen Art oder Kategorie (vgl. Roschs BASIC LEVEL TERM (1978)); Terminus SPECIES zur Bezeichnung der Subklassen; DIFFERENTIA als relevantes unterscheidendes Merkmal zur Auszeichnung einer Spezies.
- Definitionsmuster der mittelalterlichen Scholastik: Objekt wird klassifiziert durch das GENUS PROXIMUM, den unmittelbaren Oberbegriff, und die charakteristische DIFFERENTIA SPECIFICA.
- Aristotelischer Syllogismus als Schema logischen Schließens gilt als Urmodell einer Inferenzregel, besteht aus zwei Prämissen (Obersatz (1) und Untersatz (2)) sowie einer Schlussfolgerung bzw. Konklusion (3).
 - (1) Alle Menschen sind sterblich.
 - (2) Sokrates ist ein Mensch.
 - (3) Sokrates ist sterblich.

Heutzutage verstehen sich die meisten Ontologieansätze pragmatisch als sprachliche Kategorisierungen von Wissensbereichen, ohne Rückgriff auf die Metaphysik

Formal Concept Analysis (Priss)

OBJEKTE, KLASSEN und MERKMALE sind wesentliche Bausteine einer konzeptuellen Modellierung, die z.B. in einer Merkmalsmatrix zur Beschreibung realer und fiktiver Objekte mittels formaler Attribute dargestellt und in eine Gittermodellierung CONCEPT LATTICE übertragen werden kann (vgl. Priss (2006) im Kontext der „Formal Concept Analysis“ (FCA)). Ein Gitter vereint Prinzipien der Mengenlehre, Topologie und Geometrie und dient in der Mathematik zur Zerlegung eines Raumes in disjunkte Subräume.

	cartoon	real	tortoise	dog	cat	mammal
Garfield	X				X	X
Snoopy	X			X		X
Socks		X			X	X
Greyfriar's Bobby		X		X		X
Harriet		X	X			

Abbildung 1: Formale Beschreibung berühmter Tiere (Priss (2006))

Gittermodellierung (Priss (2006))

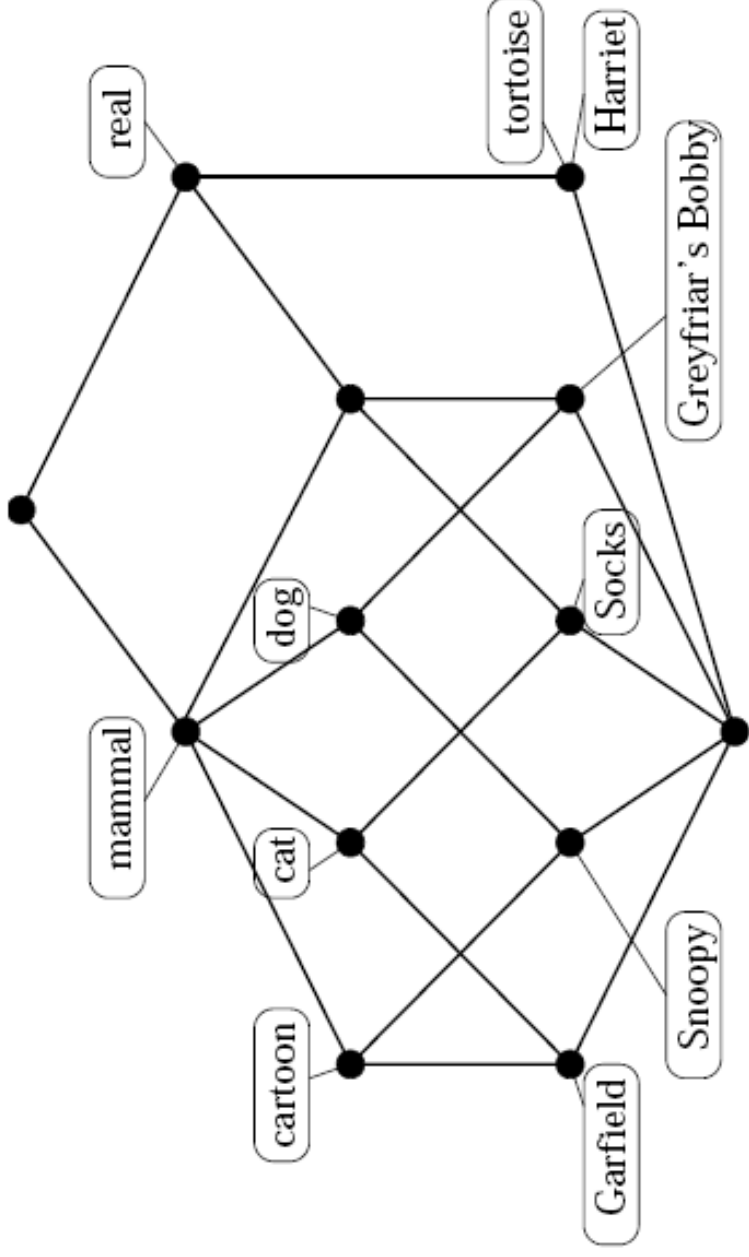


Abbildung 2: Konzeptuelle Gittermodellierung

Taxonomien und Ontologien

Formale Modellierung von Wissen: Kontinuum von relativ wenig strukturierten Listen wie z.B. Glossaren bis hin zu Ontologien. Taxonomische Gliederungen als mögliche Vorstufe zu einer Ontologie, benennen Elemente eines gegebenen Wissensbereiches und strukturieren diesen nach bestimmten Klassifikationsprinzipien, z.B. Sinnrelationen. Taxonomien haben nach Cruse (1986) folgende Eigenschaften:

- hierarchische Gliederung der klassifizierten Elemente vom Allgemeinen zum Spezifischen;
- Hyponymie zwischen Tochter- und Mutterknoten sowie Inkompatibilität zwischen Schwesterknoten sind als konstituierende Merkmale notwendig, aber nicht hinreichend (vgl. Beispiel zur Hierarchie der Tiere aus dem Kursbuch);
- Taxonomien erfordern einheitliche Strukturierungsprinzipien, die in der Hierarchie konstant durchzuhalten sind: die vertikale Relation der kettenbildenden DOMINANZ und die horizontale Relation der DIFFERENZ (d.h. eine Relation ist durch sich wechselseitig ausschließende Pfade in eine indefinite Anzahl sukzessiver Stufen durchzuhalten);
- Ebenenspezifikation in einer Taxonomie erfolgt vom Wurzelknoten (Ebene 1) aus in Entfernung zu diesem, d.h. ein Element auf der Ebene 2 ist ein Knoten abwärts vom Wurzelknoten, ein Knoten auf der Ebene n ist $n - 1$ Knoten abwärts verortet;
- natürliche Taxonomien haben häufig 5 Ebenen und enden meist auf dem wichtigsten, d.h. generischen Level: hier finden sich Vertreter der Basic Level Category wie *Tisch*, *Vogel* und *Auto*. Differenzierungen werden häufig durch Komposita realisiert.

Taxonomie nach Staab (2006)

Steffen Staabs Beispiel für eine einfache Taxonomie aus der Wissenschaftswelt mit vier Ebenen:

Taxonomy

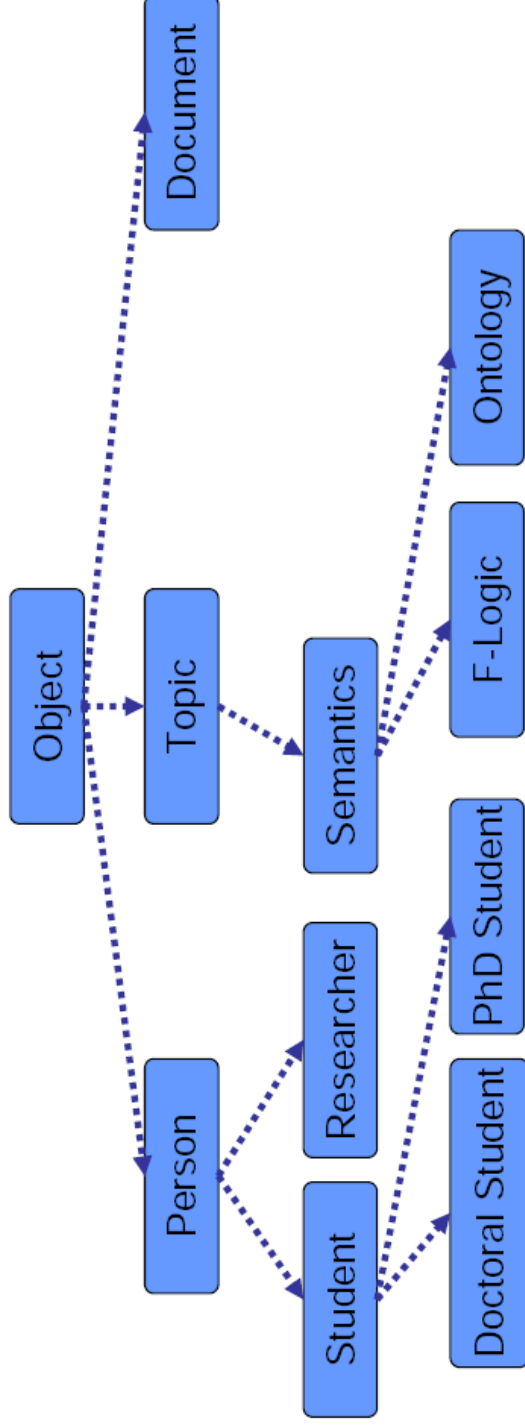


Abbildung 3: Taxonomische Gliederung einer Wissensdomäne

Thesaurus

THESAURUS: Anreicherung der taxonomischen Struktur mit sinnrelationalen Beziehungen zwischen den repräsentierten Elementen an, z.B. den expliziten Verknüpfungen der **SYNONYMIE** und **SIMILARITÄT**. Wortnetze wie GermaNet werden auch als **ONLINE-THESAURI** bezeichnet.

Thesaurus

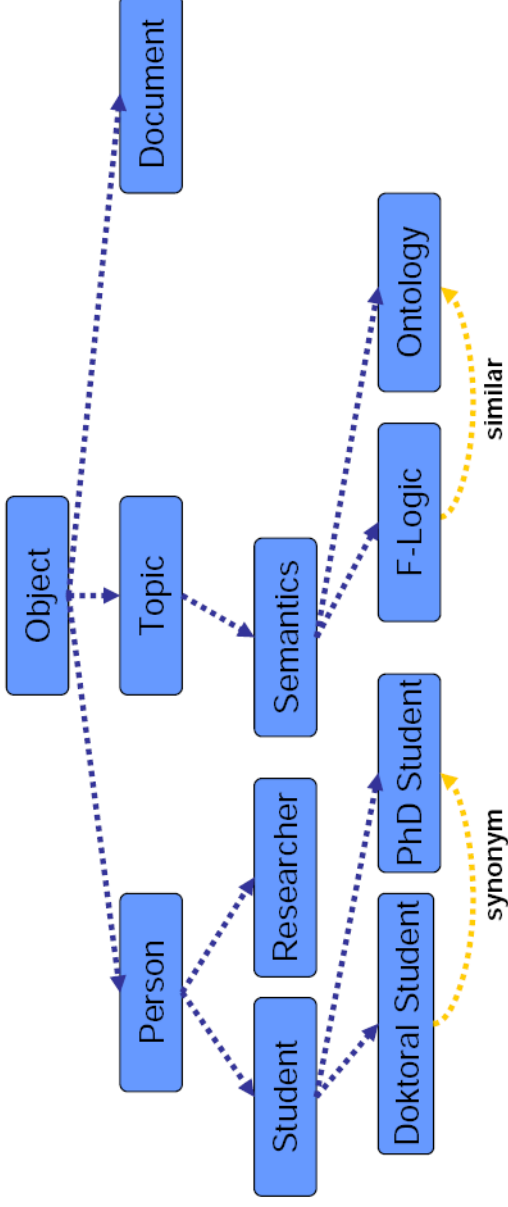


Abbildung 4: Thesaurische Gliederung der Wissensdomäne

Ontologie

- Ontologie spezifiziert die Objekte als Klassen und Instanzen in einem formalen Modell und definiert explizite Verknüpfungen zwischen den Objektklassen und Instanzen als Eigenschaftszuschreibungen
- Diese formale Basis ermöglicht Anwendung von Inferenzregeln
- Anwendung genereller logischer Regeln und Axiome, disjunkter sowie inverser Regeln sowie die Einschränkungen von Wertebereichen
- Modellierung von Ontologien zunehmend in OWL (Web Ontology Language), einem W3C-Standard, basierend auf XML und RDF(S), aber semantisch ausdrucksstärker

Ontologie nach Staab

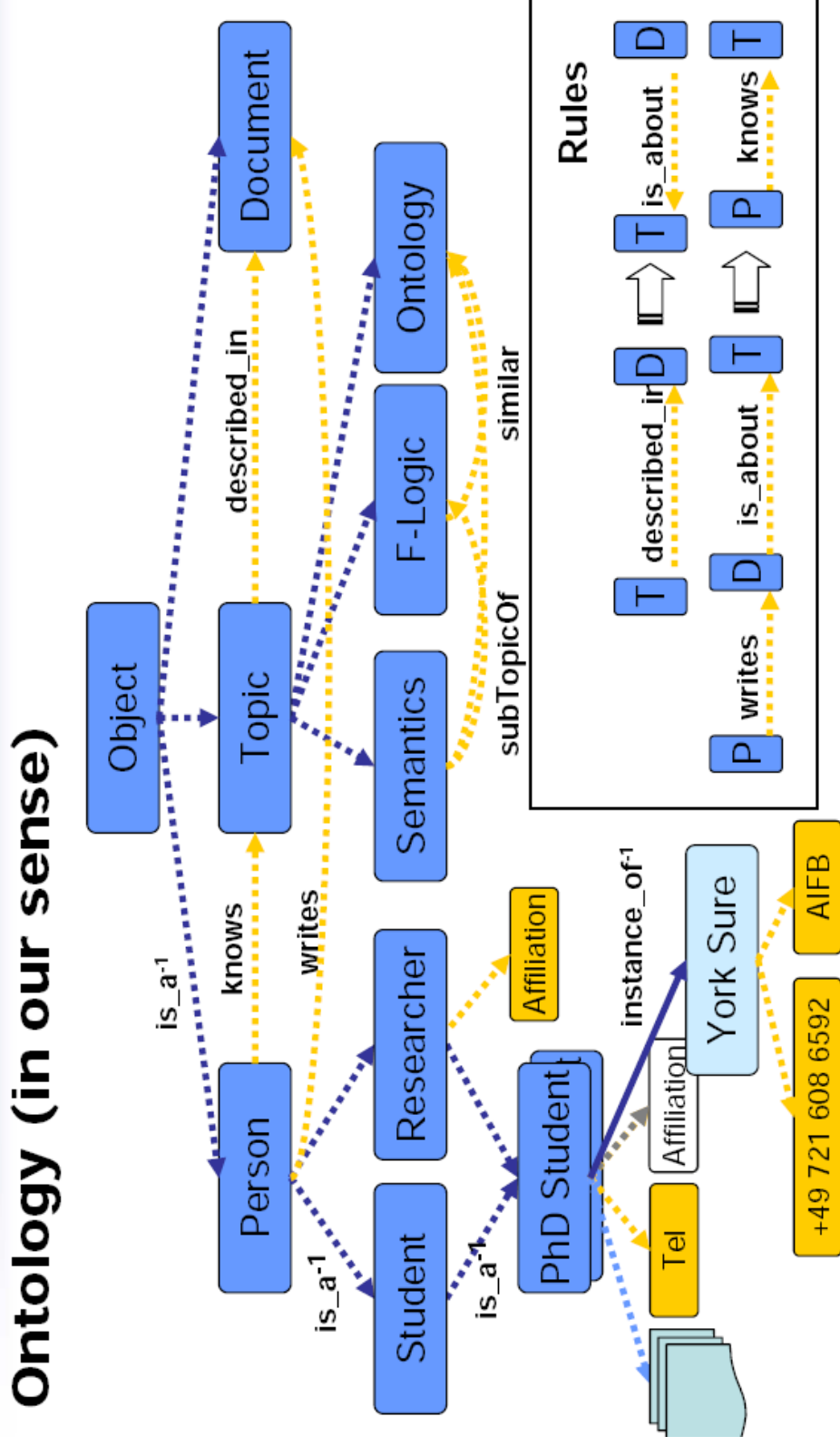


Abbildung 5: Ontologische Gliederung der Wissensdomäne

Aufbau einer Ontologie (Noy und McGuinness)

Noy und McGuinness (2001) exemplifizieren Entwicklung einer Ontologie zur Domäne **Wein**:



Festlegung der spezifischen Wissensdomäne und deren Umfang:

1. anvisierte Abdeckung der Ontologie in Bezug auf die zu modellierende Domäne und deren Spezifizität
2. Anwendungszweck der Ontologie
3. Kreis der Anwender bzw. Weiterentwickler
4. einschlägige Fragen zusammenstellen, auf welche die Ontologie Antworten bereitstellen soll: soll sowohl das Design der Ontologie leiten als auch ihrer späteren Evaluierung dienen, z.B. Fragen wie: *Welche Sorten von Rotwein gibt es? Ist Riesling ein Weißwein? Welcher Wein passt zu welchem Essen?* etc.



Einbindung existierender Ontologien zur Domäne: Zahlreiche bereits getestete Ontologien sind im Web verfügbar und über Linksammlungen zugreifbar, z.B. <http://www.ksl.stanford.edu/software/ontolLingua/>.



Zusammenstellen wichtiger Begriffe und Termini sowie ihre Eigenschaften auflisten. Welche Begriffe will man erklären oder in Aussagen einbinden, welche ihrer Eigenschaften charakterisieren? Z.B.: *wine, grape, white wine, location, color, sugar content, flavor, body* etc.

Aufbau einer Ontologie

- Erstellen der Klassenhierarchie: Die Begriffe werden nun in einer Klassenhierarchie geordnet, entweder in einem TOP-DOWN-Prozess, der vom generellsten Konzept der Hierarchie als Top-Knoten ausgehend zu den spezifischen Konzepten gelangt (von *Wine* über *White wine* und *Red Wine* zu *Syrah* und *Cabernet Sauvignon*), oder in einem BOTTOM-UP-Verfahren, das zunächst spezifische Begriffe passenden generelleren Konzepten unterordnet (etwa *Paulliac* und *Margaux* zu *Medoc*), für die wiederum eine Superklasse ermittelt wird (*Bordeaux*), oder in einer Kombination beider Verarbeitungsv Verfahren, die mit den einschlägigen Begriffen startet und diese angemessen generalisiert bzw. spezialisiert.

Domäne wine nach Noy und MacGuinness

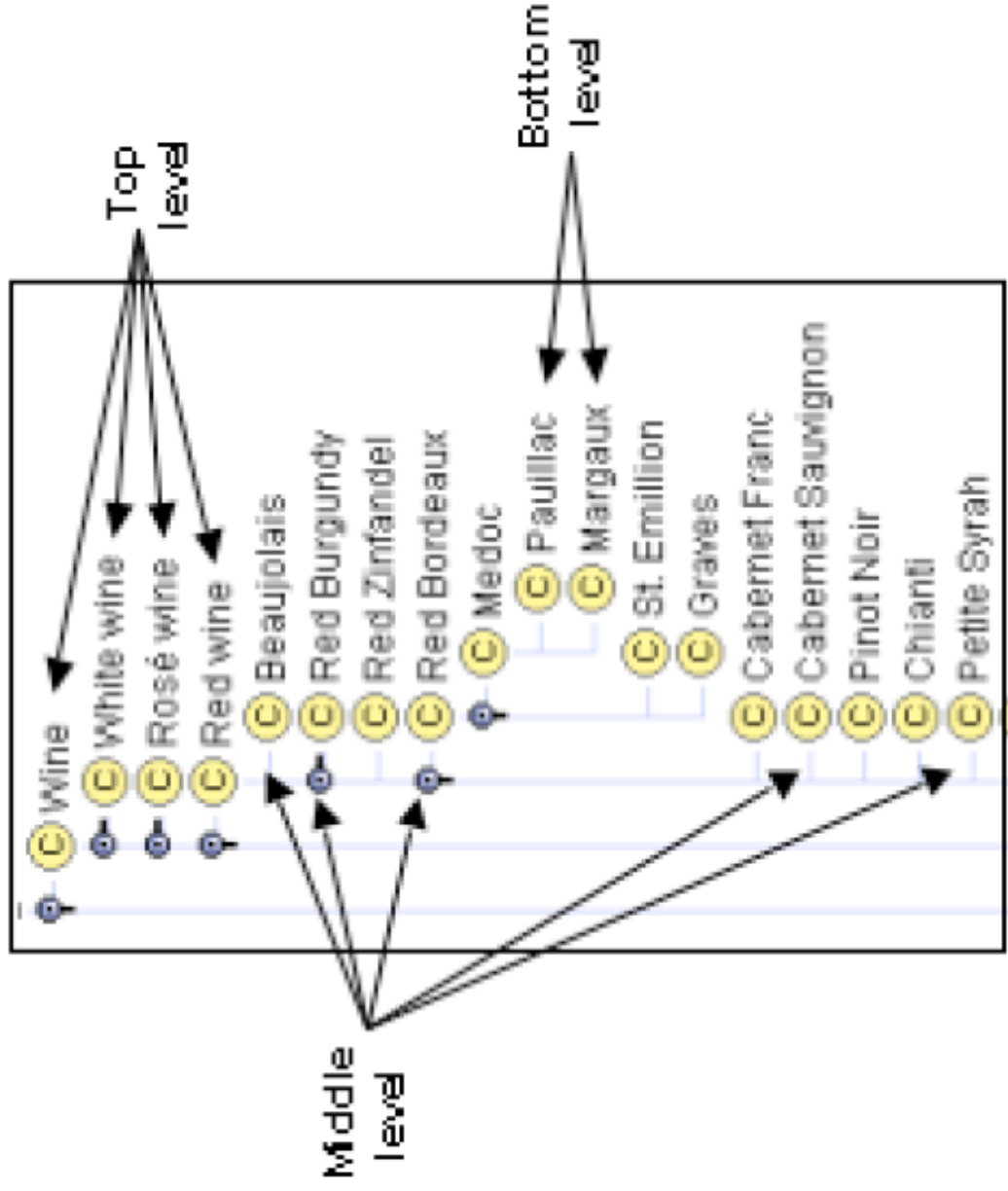


Abbildung 6: Spezifitätsebenen in der Taxonomie für wine

Klassenhierarchien

- sind IS-A Hierarchien: Ist B als Subklasse von A modelliert, so ist B „a kind of“ A (z.B. *Chardonnay* als Unterklasse von *White wine* „is a kind of“ *White wine*).
- Transitivität der Subklassenbeziehung: Wenn B eine Subklasse von A ist und C eine Subklasse von B, dann ist C eine Subklasse von A: *Chardonnay* als „kind of“ *White wine* ist auch ein „kind of“ *Wine*, wenn auch keine direkte Subklasse.
- Verändert sich die Domäne, muss die Klassenhierarchie adaptiert werden, z.B. galt *Zinfandel* immer als Subklasse von *Red wine*. Inzwischen wurde ein *White zinfandel* kreiert, dessen Farbe rosé ist. Die *Zinfandel*-Klasse muss also nun in zwei Klassen *White zinfandel* und *Red zinfandel* aufgeteilt werden und wird entsprechend entweder als *Rosé* oder als *Red wine* klassifiziert.
- Synonyme für dasselbe Konzept repräsentieren keine unterschiedlichen Klassen, sondern verschiedene Bezeichner für dieselbe Klasse, vgl. *Shrimps* und *Prawns*.
- Die Klassenhierarchie sollte keine Zyklen enthalten, d.h. nicht zulassen, dass A als Subklasse von B und B zugleich als Subklasse von A fungiert.
- Direkte Subklassen der gleichen Klasse sollten den gleichen Spezifitätsgrad aufweisen, z.B. *White wine* und *Red wine* als Unterklassen von *Wine*.
- keine genau festgelegte Anzahl direkter Subklassen zu einer gegebenen Klasse, aber die Existenz nur einer direkten Subklasse könnte auf die Unvollständigkeit der Hierarchie hinweisen. Andererseits können mehr als 12 direkte Subklassen zu einer Klasse Bedarf für intermediäre Kategorien aufzeigen.

Bestimmen der Klasseneigenschaften

- Für jede der Eigenschaften *COLOR*, *BODY*, *FLAVOR*, *LOCATION* muss die Klasse festgelegt werden, die sie beschreibt, also für *COLOR* die Klasse *wine*, und für die Eigenschaft *LOCATION* die Klasse *winery*. Die Eigenschaften werden als *SLOTS* den jeweiligen Klassen zugeordnet, und zwar auf der generellstmöglichen Ebene (z.B. beschreibt die Eigenschaft *TANNIN LEVEL* nicht die Klasse aller Weine, sondern die Klasse *Red wine* und wird genau dieser Klasse zugeordnet und an Subklassen von *Red wine* wie *Bordeaux* oder *Merlot* vererbt). Objekt-Eigenschaften in einer Ontologie können unterschiedliche Dinge bezeichnen:
 - genuine Eigenschaften wie *FLAVOR*;
 - äußere Eigenschaften wie *NAME* oder Anbauggebiet *AREA*;
 - konkrete oder abstrakte Teile eines strukturierten Objekts (z.B. als Gang einer Mahlzeit);
 - Relationen zwischen Klassenindividuen und anderen Items (z.B. *MAKER* eines *Wine*, als Repräsentant der Beziehung zwischen *Wine* und *Winery*).

Einschränkungen über Eigenschaften

- Folgende Möglichkeiten zur Einschränkung von Eigenschaften:
 - Kardinalität eines Slots, zur Festlegung der verschiedenen Werte, die er annehmen kann: Z.B. `SINGLE`, wenn nur ein Wert möglich ist oder `MULTIPLE`, wenn mehrere Werte möglich sind. Beispiel: `BODY` als Eigenschaft der Klasse *Wine* ist vom Kardinalitätstyp `SINGLE`, da ein Wein nur einen Körper haben kann: `PRODUCES` als Slot der Klasse *Winery* hat den Kardinalitätstyp `MULTIPLE`, da mehrere verschiedene Weine produziert werden.
 - Typ des Wertes eines Slots: Welche Art von Werten kann ein Slot annehmen? Z.B. den Wertetyp `STRING` für Slots wie `NAME` – der Wert ist dann eine einfache Zeichenkette – oder `ENUMERATED` für eine Liste möglicher Werte, die ein Slot wie `FLAVOR` annehmen kann: `(STRONG | MODERATE | DELICATE)`;
 - Domäne und Wertebereich (Range) eines Slots: Welche Klassen weisen die Eigenschaft auf und welche Klassen sind mögliche Werte? Z.B. die Klasse *Winery* ist die Domäne der `PRODUCES`-Eigenschaft. Als Wertebereich wird mit *Wine* die generellste Klasse angegeben, die verfügbar ist.

Erstellen der Instanzen einer Klassenhierarchie

Hierfür benötigt man individuelle Instanzen einer gegebenen Klasse, z.B.
Chateau-Morgan-Beaujolais als Instanz der Klasse *Beaujolais*:

BODY: Light

COLOR: Red

FLAVOR: Delicate

TANNIN LEVEL: Low

GRAPE: Gamay (instance of the *Wine grape class*)
MAKER: Chateau-Morgon (instance of the *Winery class*)

REGION: Beaujolais (instance of the *Wine-Region class*)

Ontologien für die Sprachverarbeitung

Es gibt eine Vielzahl von Ontologien für unterschiedliche Fachbereiche und diverse Anwendungsszenarien. Für die Linguistik sind z.B. Ontologien interessant, die helfen, Wortnetze im Bereich des Semantic Web zu nutzen. Dies kann durch eine partielle oder umfassende Anbindung an eine bestehende allgemeinsprachliche Ontologie geschehen oder durch eine direkte Konversion der Wortnetze in eine Webrepräsentationssprache wie OWL. Dazu gibt es bereits verschiedene Ansätze, vgl. van Assem et al. (2006). Verschiedene Ontologien sind Open-source verfügbar, z.B. SUMO, OpenCyc, WordNet, GUMO, DOLCE, etc.:

- Die Suggested Upper Merged Ontology SUMO (<http://www.ontologyportal.org/>) ist eine allgemeinsprachliche Ontologie, die 1500 Klassen und 4000 Axiome enthält, mit fachspezifischen Ontologien aber auf ein Vielfaches erweitert werden kann. Die generischen Konzepte des WordNet sind mit SUMO und dessen Top-Level-Struktur verlinkt, was für viele sprachtechnologische Anwendungen vorteilhaft ist.
- Das DOLCE/OnoWordNet-Projekt (<http://www.loa-cnr.it/DOLCE.html>) ist noch stärker auf die Alignierung mit WordNet konzipiert. Im Mittelpunkt steht eine möglichst breite Abdeckung der WordNet-Konzepte als Klassen und Instanzen der DOLCE-Ontologie, wobei formale Inkonsistenzen des WordNet überprüft und bereinigt werden.

Ontologien für die Sprachverarbeitung

- OpenCyc (www.openencyc.org/) hingegen hat sich eine breitere Abdeckung, nämlich die Erfassung des gesamten Weltwissens zum Ziel gesetzt. Neben Alltagswissen ist auch umfangreiches Fachwissen enthalten, zur breit angelegten Unterstützung maschinellen Inferierens. Mittlerweile sind mehr als 100 000 Begriffe mit mehr als 1 000 000 Eigenschaften und Fakten repräsentiert.
- Bei der General User Model Ontology GUMO (<http://www.ubisworld.org/>) steht die Modellierung durch den Anwender im Vordergrund, der sein Modell anhand verschiedener Dimensionen wie RAUM, ZEIT, FÄHIGKEITEN, EREIGNISSE und GEGENSTÄNDE entwickeln kann. GUMO ist in OWL entwickelt und stellt das Vokabular für die verschiedenen Bereiche zusammen, ohne Axiome und Restriktionen zu determinieren. GUMO ist eine interessante Plattform zur Simulation realer (Teil-)Welten, wie z.B. Geschäfte, Museen und Flughäfen.

Auch in zukünftigen Forschungen werden Ontologien eine wichtige Rolle spielen, z.B. hinsichtlich ihrer Verknüpfbarkeit mit lexikalischen Ressourcen.