**Pascal Christoph**

# Paradigmenbildung in einem selbstlernenden System

**Magisterarbeit**

Examicus

Universität zu Köln
Philosophische Fakultät
Institut für Linguistik, Abt. Sprachliche Informationsverarbeitung

# Paradigmenbildung in einem selbstlernenden System

Version 0.41

Magisterarbeit

von
Pascal Christoph

9. Semester
Sprachliche Informationsverarbeitung
Phonetik
Allgemeine Sprachwissenschaft

Köln, im März 2006

# Inhaltsverzeichnis

# 1. Einleitung

*»The Ultimate information system will work so well that finding information is as easy as remembering it.«*

*Gregory B. Newby*

In der sogenannten *Informationsgesellschaft* lebend bedarf es Mittel und Techniken, die Inhalte, Informationen und Dokumente verschiedenster Art zu organisieren, um mehr als nur zufälligen Zugriff auf das Gewünschte zu erhalten.[1] Mittlerweile hält nicht mehr nur das weltweite Internet eine unglaubliche Fülle an Dokumenten bereit, auch der heimische Rechner ist längst zur Bibliothek geworden, in der sich zehn Jahre Tageszeitung neben Teilen der Gutenbergbibliothek[2], der Wikipedia[3] und einer Menge anderer Bücher, Aufsätze etc. befinden. Der massive Zuwachs an digitalen Dokumenten lässt für deren Organisation keine manuelle, sondern nur eine automatische Methode zu. Dabei sind zumindest für die textuellen Dokumente neben den bestehenden Werkzeugen von dem Sprachlichen Informationsverarbeiter weitere Werkzeuge entwickelbar, wie das in der vorliegenden Arbeit entwickelte Programm beispielhaft aufzeigt. Eine Möglichkeit läge darin, die Meta-Tags[4] in HTML-Dokumenten um Keywords zu erweitern, die nicht explizit Bestandteile des originären Dokuments sind. Ein mögliches Verfahren dazu ist die Generierung von Paradigmen, denn Worte sind im gleichen Paradigma wenn sie „gegeneinander austauschbar“[5] sind, also z.B. die Wörter *Orange* und *Apfelsine*[6] oder die Wörter *Obst* und z.B. *Orange*[7]. Die auf diese Weise ausgezeichneten Dokumente, in denen etwa im eigentlichen Text stets nur von *Apfelsine* die Rede ist, würden auch dann noch durch eine Suchmaschine auffindbar sein, wenn der Benutzer als Suchbegriff *Orange* eingegeben hätte. Für die Analyse und maschinelle Erzeugung von Wissen ist demnach

1 Mit zunehmender Dokumentenzahl steigt die Unwahrscheinlichkeit des Rechercheerfolges. Ein ausführlicher Artikel zum Thema findet sich unter:

http://de.wikipedia.org/wiki/Dokumentenmanagement (letzter Zugriff: 21.02.2006)

2 http://www.gutenberg.org/ (letzter Zugriff: 21.02.2006)

3 http://de.wikipedia.org/wiki/Hauptseite (letzter Zugriff: 21.02.2006)

4 *Meta-Tags* sind versteckte Elemente auf einer Webseite. Sie enthalten Metadaten über das betreffende Dokument.

5 Diese Definition gilt für die Domäne Linguistik. Es gibt andere Lesarten von *Paradigma*.

6 Der Fachterminus einer solchen Relation lautet *Synonym***.** Dabei hat das Beispiel nur für „die Nordhälfte Deutschlands Gültigkeit und ist in Österreich und der Deutschschweiz als Teutonismus markiert. In Bayern würde der Gebrauch des Wortes Apfelsine einen "Zugereisten" oder Urlauber kennzeichnen.“ Vgl. http://de.wikipedia.org/wiki/Synonym (letzter Zugriff: 08. 01. 2006)

7 Der Fachterminus einer solchen Relation lautet *Hyperonym* (vgl. Kapitel 2.12)**.**

eine Akzentverschiebung in Richtung Semantik[8] erfoderlich.

In jüngster Zeit wird die automatisierte Paradigmenbildung als wichtiger Bestandteil zur Informationsorganisation fokussiert.[9] Dieser Umstand ist insbesondere der taxonomischen Eigenschaft von Paradigmen geschuldet: sowohl Elemente innerhalb eines Paradigmas, als auch Paradigmen als Ganzes (vgl. dazu Schwiebert 2004:9 f.), stehen immer in Relation zueinander.[10] Das XML-Magazin veröffentlichte zu dem Thema Taxonomie und Topic Maps einen Aufsatz von Thomas Bandholtz. Hier beobachtet Bandholtz (2002):

> Internationale Firmen (wie z.B. Accenture, BP, HP, IBM, Microsoft, Nokia, Royal Dutch/Shell, Schlumberger, Siemens, Toyota oder Xerox ) wetteifern im Aufbau ihrer Master Classification und weisen terminologisch versierten Mitarbeitern eine Rolle als Taxonomist zu.

Auch in mittelständischen Firmen steigt der Bedarf an Organisation digitaler Dokumente.[11]

Ferner findet sich in dem Artikel von Bandholtz ein Querverweis auf den in *The Bulletin: Seybold News & Views On Electronic Publishing* veröffentlichten Text von Luke Cavanagh, in dem betont wird, dass Taxonomie Managemet eine Schlüsselposition im Content Management erhält. Cavanagh (2002) stellt fest: „The categorization software business is developing as we speak, and the software being created may well be the next big must-buy item in your organization."

Wie der Titel der vorliegeden Arbeit „Paradigmenbildung in einem selbstlernenden System" impliziert, ist zentrales Ziel dieser Arbeit die Entwicklung einer Software, die automatisch Paradigmen auf der Grundlage eines Korpus generiert. Das dazu notwendige linguistische Fundament wird in Kapitel 2 dargestellt. Im Anschluss an Jürgen Rolshovens (2002:3) Feststellung „Die Codierung in der parole ist sehr unvollständig."[12], wird in diesem Kapitel u.a. den Fragen nachgegangen, inwiefern sich diese Unvollständigkeit auf die richtige Bestimmung von Paradigmen auswirkt. Wo liegen die Grenzen der Computabilität von Paradigmen bzw. welche Schwierigkeiten gibt es? Bereits die

8 Die Semantik ist ein Synonym zu *Bedeutungslehre*.

9 Im deutschsprachigen Raum wurde das Projekt des Leipziger Wortschatz entwickelt. Vgl.: http://wortschatz.uni-leipzig.de/ (letzter Zugriff: 21.02.2006)

10 Auch wenn Paradigmen *per se* keine wissenschaftliche Klassifizierung der außersprachlichen Realität darstellen, so organisieren sie doch linguistische Klassifikationen.

11 Vgl. auch: http://www.knowledgebusiness.com/knowledgebusiness/ (letzter Zugriff: 21.02.2006)

12 Der Begriff *parole* ist von Ferdinand de Saussure geprägt und bezeichnet den tatsächlichen Sprachgebrauch, also z.B. Sätze, im Gegensatz zur *langue*, die das allgemeine Regelwerk von Sprache beschreibt, wie etwa die Grammatik.

Geschichte der Sprachwissenschaft liefert darauf eine erste Antwort. Dies aufzuzeigen, ist u.a. die Aufgabe des 2. Kapitels. Zusätzlich liefert das Kapitel 3.3 mit einer Diskussion über die technischen Grenzen und Probleme praxisnahe Einsicht in das weite Feld der strukturellen Semantik.

Es besteht ein weitverbreitetes Interesse der Wirtschaft sowie öffentlicher Institutionen an der Bereitstellung einfach zu bedienender Schnittstellen zwischen Mensch und Maschine. Aus diesem Grund war der Einsatz entsprechender Softwaretechnologien sowie die Konzeption und Programmierung einer „niedrigschwelligen" Schnittstelle ein zentrales Anliegen vorliegender Arbeit. Die Verwendung des Programms soll für den Benutzer möglichst unkompliziert sein.

Das Ziel der Arbeit besteht folglich in der Erzeugung von *deklarativem Wissen* mittels *funktionalen Wissens.*[13] Die maschinelle Analyse eines beliebigen Korpus[14] erzeugt durch einen Algorithmus[15] Daten[16], die in einer Datenbank gespeichert werden.[17] Das auf diese Art generierte sprachliche Wissen, die Zuordnung von Wörtern zu einem Wortparadigma, sollte im Idealfall dem intuitiven Wissen eines Muttersprachlers nicht widersprechen.[18] Ein wichtiges Anliegen ist das in der Datenbank persistierte deklarative Wissen jedem Anwender frei zur Verfügung zu stellen und den Zugriff auf diese Daten möglichst einfach[19] zu halten. Im Kapitel 3 wird das aus der theoretischen Vorarbeit entstandene Programm *PaGe*[20] vorgestellt und seine Handhabung erläutert.

Im letzten Kapitel werden zusammenfassend die Ergebnisse der Arbeit dargelegt sowie offene Fragestellungen und Wünsche aufgezeigt.

Vor dem Übergang zum nächsten Kapitel wird an dieser Stelle die in der Arbeit verwendete Notation erläutert:

*Kursiver Text* soll das Rezipieren erleichtern. Außer in Kapiteln vorangestellten Zitaten

13 Qualitativ unterscheidet sich deklaratives Wissen vom Output funktionalen Wissens lediglich durch die Latenzzeit. *Latenzzeit* beschreibt den Zeitraum zwischen einer Aktion (hier: der Anfrage, in welchem Paradigma ein Wort liegt) und dem Eintreten einer Reaktion (hier: die Antwort der Zuordnung des Wortes zu einem Wortparadigma).

14 Einschränkungen hinsichtlich des Aufbaus des Korpus werden in Kapitel 3.2 erläutert.

15 Der *Algorithmus*, also ein exakt definierter Handlungsablauf, stellt das funktionale Wissen dar.

16 Die *Daten* stehen für das deklarative Wissen. Sie sind entweder axiomatisch oder (wie im vorliegenden Fall) Produkte der Anwendung anderer Daten auf funktionales Wissens.

17 Siehe hierzu Kapitel 3 und Kapitel 4

18 Ergebnisse des PaGe werden in Kapitel 4 diskutiert.

19 Siehe hierzu den Exkurs in Kapitel 3.1 zur Barrierefreiheit

20 PaGe ist ein aus den beiden Anfangsbuchstaben der Worte *Paradigmen* und *Generator* gebildetes Akronym.