

Struktur der Magisterarbeit Pascal Christophs V1.0

Möglicher Titel:

"Wortzuordnung zu linguistischem Paradigma mittels selbstlernenden Systems. Eine J2EE Realisierung in der Domäne Quantitative Linguistik."

Theoretisierung des linguistischen Paradigmbegriffs:

- Wikipedia: "Sammlung von (auf vertikaler Ebene) austauschbarer Zeichen (Elemente) der selben (Wort)Kategorie": 1.nur grammatisch (Restriktionsbedingungen) , 2.grammatisch *und* semantisch
- Metzlers Sprachlexikon
- Saussure: assoziative Beziehungen. In: Grundfragen der allgemeinen Sprachwissenschaft
- John Lyons: paradigmatische Relationen. In: Einführung in die moderne Linguistik

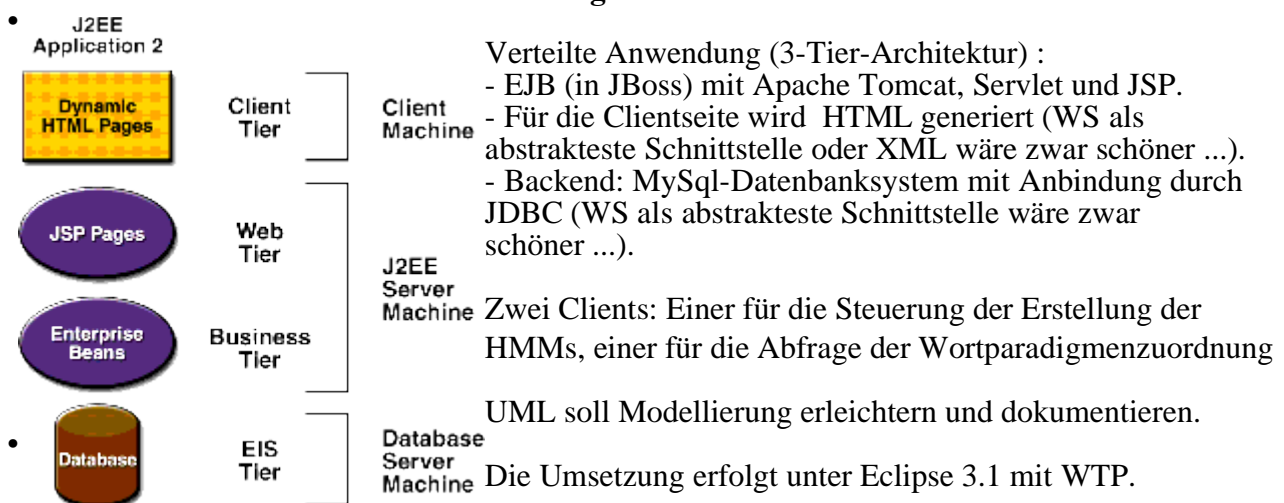
Probleme der Repräsentation des Paradigmbegriffs im Kontext Semantik (z.B. Polyseme):

- Lothar Schmidt (Hsg.): Wortfeldforschung
- George Lakoff: Women, Fire, and Dangerous Things
- Manfred Spitzer: Geist im Netz
- Howard Gardner: Dem Denken auf der Spur

Beschreibungsadequatheit <-Deduktion <- Algorithmen auf Grundlage der Mengenlehre <- Beobachtungsadäquatheit <- Markov Modellen auf Grundlage der Auswertung von Korpora:

- HMM sowohl für rein grammatikalische als auch semantische paradigmatische Relationen
- Auswertung schlichten ASCII's auf Webservern, später Auswertung z.B. auch des Tamino-Korpus'
- Persistierung der Modelldaten auch von >20 GB Daten in mySQL DBS
- Literatur u.a. :
 - Böckenbauer, Bongartz: Algorithmische Grundlagen der Bioinformatik
 - Gernot A. Fink: Mustererkennung mit Markov-Modellen

Dokumentation der technischen Umsetzung:



Grafikentnahme aus dem J2EE 1.4 Tutorial von Sun

Fragestellungen/Ziele:

- Verlagerung von sprachstrukturellen Informationen in das Lexikon. Bei Ambiguitäten Zugriff auf Transitionswahrscheinlichkeiten (Hidden-Markov-Modelle)
- Gelingt die Paradigmazuordnung im Kontext Semantik vollkommen/eingeschränkt/nicht?
- Markov Modelle sollen Suffixbäume verarbeiten können
- Daten für die Erstellung der Markov Modelle sollen sowohl von links als auch von rechts eingelesen werden können
- Es sollen Markov Modelle höherer Ordnung möglich sein
- Datenreduktion durch "dynamische n-Dimensionale Matrizen": nur tatsächliche Okkurenzen werden gespeichert
- Klärung der theoretischen Grenzen des Systems

Fakultative Fragestellungen/Ziele, die zu untersuchen sich (vielleicht) lohnen:

- Anwendung des Selbstlernenden Systems auf verschiedene (nicht-)indogermanischen Sprachen, z.B. dem Inuit
- Lassen sich durch verschiedene Korpora Domänenspezifika finden (z.B. Biologie, Religion ...) ? Bringt eine Einteilung in Subdomänen (z.B. Biologie -> Genetik, Religion -> Buddhismus) weitere Erkenntnisse ? (Auch Autorenspezifika, Epochenspezifika usw. ließen sich untersuchen ...)
- Gibt es Paradigmen im Kontext Prosodie und was lehrt eine Interpretation? GToBI
Korpusauswertung.
- Gibt es Paradigmen im Kontext Silbe und was lehrt eine Interpretation? Silbenannotierte
Korpusauswertung.
- Gibt es Paradigmen im Kontext Phrase und was lehrt eine Interpretation? Phrasenannotierte
Korpusauswertung.
- Gibt es Paradigmen im Kontext Satz ("Topic-Domäne") und was lehrt eine Interpretation?
Satztopicannotierte Korpusauswertung.
- Bringt die Verknüpfung dieser verschiedenen Ebenen Vorteile bei der Disambiguierung ?
- Visualisierung der Modelle mit OpenGL-Vektorgrafik als 3-dimensionale Wortfelder wäre schön
(4-dimensional (3-dimensionale Animation): z.B. Sichtbarmachung diachronischen Wandels).